



Published in final edited form as:

Neuron. 2015 July 15; 87(2): 411–423. doi:10.1016/j.neuron.2015.06.033.

How can single sensory neurons predict behavior?

Xaq Pitkow^{1,2,*}, Sheng Liu¹, Dora E. Angelaki^{1,2}, Gregory C. DeAngelis³, and Alex Pouget^{3,4}

¹Baylor College of Medicine, Department of Neuroscience, One Baylor Plaza, Houston TX 77030 USA ²Rice University, Department of Electrical and Computer Engineering, 6100 Main MS-366, Houston TX 77005 USA ³University of Rochester, Department of Brain and Cognitive Sciences, 358 Meliora Hall, Rochester NY 14607 USA ⁴University de Genève, Dept of Neuroscience, 1 rue Michel-Servet, 1211 Geneva 4 Switzerland

Summary

Single sensory neurons can be surprisingly predictive of behavior in discrimination tasks. We propose this is possible because sensory information extracted from neural populations is severely restricted, either by near-optimal decoding of a population with information-limiting correlations or suboptimal decoding that is blind to correlations. These have different consequences for choice correlations, the correlations between neural responses and behavioral choices. In the vestibular and cerebellar nuclei and the dorsal medial superior temporal area, we found that choice correlations during heading discrimination are consistent with near-optimal decoding of neuronal responses corrupted by information-limiting correlations. In the ventral intraparietal area, the choice correlations are also consistent with the presence of information-limiting correlations, but this area does not appear to influence behavior although the choice correlations are particularly large. These findings demonstrate how choice correlations can be used to assess the efficiency of the downstream read-out and detect the presence of information-limiting correlations.

Introduction

Individual sensory neurons in the brain are often predictive of animals' choices in simple perceptual decision-making tasks. It is said that these neurons have a significant choice probability. This remarkable fact has been demonstrated in numerous tasks and brain areas, including those dedicated to sensing visual motion (Britten, Newsome et al. 1996), depth (Uka and DeAngelis 2004, Nienborg and Cumming 2007), and self-motion (Gu, Angelaki et al. 2008, Fetsch, Pouget et al. 2012, Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). Many of these cells have neural thresholds, which quantify sensitivity to stimulus variations, that are not much greater than psychophysical thresholds (Cohen and Newsome 2009). It is therefore puzzling why pooling these signals does not predict sensitivity much greater than that exhibited by behavior. Perhaps the brain merely selects a small subset of neurons to inform its decisions (Tolhurst, Movshon et al. 1983, Ghose and Harrison 2009) — but then how could experiments so frequently encounter these extremely rare neurons that influence behavior? A proposed explanation for these puzzling observations was that response

*Corresponding author: xaq@rice.edu.

variability is correlated across neurons (Zohary, Shadlen et al. 1994): even with very weak correlated noise between pairs of neurons, the total information content of a neural population may saturate to a finite value as the number of neurons increases, such that optimally pooling more responses cannot improve behavioral sensitivity. Additionally, neurons are not only correlated with each other but also with the pooled signal that presumably drives the perceptual decision, which would generate high choice probabilities.

This solution (Zohary, Shadlen et al. 1994) was established for a very simplified model of neural responses, correlations, and decoding. Subsequent studies relaxed some of these simplifications, and found consistent results for broad correlations in neural populations tuned to a one-dimensional stimulus (Sompolinsky, Yoon et al. 2001). However, it was suggested that diversity in the amplitude and width of neural tuning curves would change the picture (Abbott and Dayan 1999), and later calculations demonstrated that weak noise correlations do *not* limit information in heterogeneous neural populations: information continues to increase linearly with the number of neurons (Shamir and Sompolinsky 2006, Ecker, Berens et al. 2011). We say that such a population has ‘extensive information’. If correct, this would imply that correlated noise cannot explain the frequent occurrence of significant choice probabilities, for the following reason: in optimally decoded populations with extensive information, each neuron provides a tiny contribution, inversely proportional to the size of the neural pool, towards the perceptual decision. This prediction is at odds with observed choice probabilities and ratios of neural to psychophysical thresholds.

Perhaps the neural population contains vast amounts of information, but it is not all used in perception. There are many forms of such suboptimal decoding that misuse neural signals. We will show that suboptimal decoding could indeed explain both why behavioral thresholds are barely better than single neuron thresholds and why choice probabilities are so large and common.

A second explanation of these phenomena does not rely on suboptimal decoding but instead blames a subtle form of neural noise correlations (Moreno-Bote, Beck et al. 2014), called differential correlations, that limit the information contained in a population code. These information-limiting noise correlations cause massive redundancy between neurons, which restricts behavioral thresholds to be not much better than individual neural thresholds. We show that this explanation also predicts many neurons with high choice probabilities.

Thus both suboptimal decoding and information-limiting noise correlations could explain these two puzzling phenomena. Which is the correct explanation? We derive quantitative consequences of each hypothesis in order to understand the nature of neural population codes. We test these consequences in various brain areas that are responsive to vestibular signals and are activated during a heading discrimination task. We find that most of the data are more consistent with near-optimal decoding of neural responses with information-limiting noise correlations.

Results

During a vestibular heading discrimination task, animals were presented with a movement stimulus s , specifically, translation by a motorized platform within the horizontal plane (Supplementary Figure S1). S/he must use the responses of neurons tuned to the vestibular stimulus s in order to estimate a direction of motion \hat{s} , and to discriminate whether that heading is slightly leftward or rightward of some reference heading s_0 , which was straight forward in our task. The heading estimate is generated by pooling responses \mathbf{r} of neurons in some way. In the brain areas from which we recorded, neurons are tuned to heading, with average responses, $f_k(s)$, that are characterized by a few parameters for each neuron k , including its preferred heading s_k (Figure 1A, Methods).

The ability of a single neuron to discriminate between similar headings is generally greatest when the tuning curve has a steep slope $f'_k = df_k/ds$ near the reference stimulus, such that the mean response changes substantially with small variations in heading. However, neural responses are noisy, varying from trial to trial even when the stimulus is the same.

Consequently, discriminability decreases for larger response variance σ_k^2 . We define a discrimination threshold θ_k for each neuron as the signal change required to exceed one standard deviation of noise, $\theta_k = \sigma_k / f'_k$.

An animal may estimate the stimulus more reliably than single neurons by pooling signals appropriately across a population of neurons. To understand the information content of the population, it is helpful to visualize how the vector of mean responses traces out a curve, $\mathbf{f}(s)$, in the N -dimensional neural response space as a function of the stimulus s (Figure 1B). For the fine discrimination tasks we examine here, the tested stimulus range around the reference is sufficiently narrow that the mean neural responses depend nearly linearly on the stimulus (Gu, Angelaki et al. 2008), thus lying close to the tangent vector \mathbf{f}' . Over such a narrow stimulus range, evidence from other systems suggests that most of the information can be extracted near-optimally by a linear decoder (Ma, Beck et al. 2006, Graf, Kohn et al. 2011, Berens, Ecker et al. 2012). We therefore model the animal's estimate \hat{s} as a linear weighting of all neural responses \mathbf{r} in a population (Figure 1C), according to

$$\hat{s} = \mathbf{w}^T (\mathbf{r} - \mathbf{f}(s_0)) \quad (1)$$

This linear decoding can be viewed as a projection of the N -dimensional responses onto a single dimension defined by the weight vector \mathbf{w} (Figure 1B).

Noise in the neural population generates a cloud of possible responses around the mean response. The covariance Σ of this high-dimensional response variability can be visualized as an ellipse centered on the mean response (Figure 1B). Among the many dimensions of this noise, only noise along the decoding direction \mathbf{w} generates variability in the estimate \hat{s} ; the remaining variability in the orthogonal directions has no effect on the estimate. We can define a discrimination threshold θ for the decoded estimate just like we do for single neurons, as the amount that the stimulus must change for the estimate to exceed one

standard deviation of noise in the estimate, $\sigma_{\hat{s}}$ (see Methods). This threshold is represented by the width of a gaussian distribution centered on the mean estimate (Figure 1B).

Animals can be trained to give largely unbiased reports in this task (Methods), which means that on average the animal has an accurate estimate. Although behavior should benefit from combining information from many neurons, behavioral discrimination thresholds are not substantially better than thresholds for the best single neurons (Gu, Angelaki et al. 2008, Cohen and Newsome 2009). Is this because the responses of these neurons are correlated, such they don't provide independent information? Or is the brain using their information poorly? We can distinguish these possibilities by looking across trials at the relationship between neural responses and perceptual reports.

This relationship is typically quantified by the ‘choice probability’, which is the probability that a neural response associated with one behavioral choice is greater than a neural response associated with the other possible choice (Britten, Shadlen et al. 1992). As derived by Haefner et al. (2013), choice probability is influenced both by the neural correlations (via the noise covariance matrix Σ) and by the decoding weights \mathbf{w} (Methods). Here, we measure the relationship between neuron and behavior by computing the ‘choice correlation’, the Pearson correlation coefficient C_k between the response r_k of neuron k and the estimated stimulus \hat{s} , $C_k = \text{Corr}(\hat{s}, r_k)$. This quantity has a simple, nearly affine relationship to choice probability (see Haefner, Gerwinn et al. 2013 and Methods) and an exact proportionality to the correlation between r_k and the binary choice $\text{sgn}(\hat{s})$ (Methods), but it is conceptual simpler and mathematically more convenient. Below we extend the important results of (Haefner, Gerwinn et al. 2013) to analyze choice correlation under conditions of information-limiting correlations, suboptimal decoding, or both.

Consequences of suboptimal decoding on choice correlations

One possible account of high choice correlations is suboptimal decoding. Perhaps the information encoded by areas representing heading is indeed extensive, growing in proportion to the number of neurons, but the downstream neural circuits fail to extract information efficiently. This could predict psychophysical thresholds that are not vastly better than the typical single-neuron discrimination threshold. Would this mechanism also produce high choice correlations?

We examined a family of suboptimal decoders that are blind to the patterns of correlated fluctuations present in the neural population. These are decoders that are based solely on the signal strength in individual neurons, and do not take into account the correlations between neurons. For instance, one commonly used decoder of this type is known as a ‘factorial decoder’. The factorial decoder assumes that all neurons are independent, so the probability of a population response factorizes over neurons. This decoder simply weights each response according to the individual neural sensitivities. This is a maximum likelihood decoder only when the neurons are truly independent (Földiák 1993, Sanger 1996, Liu, Gu et al. 2013), an assumption which is generally violated in the brain. Nonetheless, in some circumstances, the factorial decoder is nearly optimal, despite unfaithfully neglecting correlations (Wu, Nakahara et al. 2001). In other circumstances, such as in the presence of tuning curve diversity and smoothly varying noise correlation coefficients (Shamir and

Sompolinsky 2006, Ecker, Berens et al. 2011), this correlation-blind decoder is extremely suboptimal and throws away almost all of the information. More specifically, while the information in the population may be extensive, the information recovered by the factorized decoder is not. Instead, the information saturates to a finite value, producing a high behavioral threshold (Supplemental Information, also see (Shamir and Sompolinsky 2006) for the similar population-vector decoder). In this case, instead of cancelling the large noise fluctuations shared by many neurons, the suboptimal, correlation-blind decoder would preserve them and behavior would be largely driven by this irrelevant noise. Many neurons would be strongly correlated with behavior because they share the strong correlated noise that drives it. This could explain the prevalence of high choice correlations.

To be more quantitative, we model the noise correlation coefficient matrix R in accordance with recent experimental studies (Cohen and Kohn 2011, Liu, Gu et al. 2013). Specifically, we assume that noise correlations are proportional to signal correlations on average, with proportionality constant c_0 (Eq. 8), but with substantial heterogeneity around this average trend (Eq. 10) (Figure 1D). Such noise does not limit the information content of a heterogeneous population (Shamir and Sompolinsky 2006, Ecker, Berens et al. 2011). Yet the resultant extensive information can only be extracted if these noise correlations are cancelled by appropriate weighting of the neurons, which is not the case for suboptimal, correlation-blind decoders. Under the faulty assumption of independent neurons, the suboptimal factorial decoder uses weights given by $\mathbf{w} \propto F^{-1}\mathbf{f}'$, where F is a diagonal matrix with the correct variances but no correlations and $\mathbf{f}'=d\mathbf{f}/ds$ represents the derivative of the tuning curves evaluated at the reference stimulus. When the neurons are correlated according to Eqs. 8–10 (Methods, Figure 1D), we show that the choice correlations for a correlation-blind decoder are well-approximated by

$$C_k^{\text{cb}} \approx \sqrt{c_0} |\sin s_k| \quad (2)$$

where s_k is the preferred stimulus of the neuron relative to the reference stimulus $s_0=0$, and c_0 is the proportionality constant between signal correlations and noise correlations (see Methods and Supplemental Material). This relationship (Eq. 2) reflects not the individual neural sensitivities, but rather the structure of the broad noise correlations that are not removed by the decoder. When correlated noise is not removed, shared fluctuations dominate the animal's resulting choice, and choice correlation depends on the relationship between the neural response and the correlated noise.

We can understand Eq. 2 by examining both the suboptimal decoding weights and the noise correlation structure. For the family of suboptimal, correlation-blind decoders we consider, the weight assigned to each neuron tends to have the same sign as its tuning slope at the reference stimulus. Recall that the noise correlations (Eqs. 8–10) are strongest for neurons with similar preferred stimuli, so neurons with preferred directions near the reference stimulus will be positively correlated with other neurons whose preferred stimuli are also near the reference stimulus — but many of these preferences will fall on the other side of the reference and thus have opposite signs of decoding weights. These neurons' response fluctuations are therefore largely cancelled by the decoder, and, thus, cells with preferences

near the reference will have weak correlations with the output (Supplemental Figure S2A). In contrast, neurons with preferred stimuli farther from the reference will be correlated with many other neurons weighted with the same sign, and not with neurons on the other side of the reference that are weighted by the opposite sign (Supplemental Figure S2B). As a result, there is little cancellation of noise within this group and their collective fluctuations successfully drive the output. This explains why Eq. 2 shows choice correlations determined by the distance between each cell's preferred heading and the reference stimulus (Eqs. 8–9, Supplemental Figure S2C, Methods). Interestingly, Eq. 2 holds even in the presence of substantial variability in correlations, baseline neural activity, amplitude heterogeneity, and tuning curve width (Figure 2), as well as in the presence of random deviations (Eq. 10, Methods) around the trend relating noise correlations to signal correlations (Eq. 8–9, Methods).

The same result holds for other correlation-blind decoders that weight neurons solely by some (possibly stochastic) function of their individual sensitivities (Supplemental Figure S3A), and even when the relationship between noise correlations and signal correlations is very weak (Supplemental Figure S3B).

Suboptimal decoding could also account for a wide range of empirical observations regarding the average strength of choice correlations. According to Eq. 2, choice correlations depend on the overall correlation scale given by the proportionality constant, c_0 , between signal and noise correlations (Methods, Eq. 8), which is typically in the range of 0.1 to 0.5 (Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). Thus, a steeper slope in the relationship between noise and signal correlations would lead to greater choice correlations in this regime, and some empirical studies have reported results that are consistent with this prediction of suboptimal decoding (Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). In this situation, choice correlations are readily distinguishable from chance and do not decrease with the number of neurons or the overall information content in the population.

Consequences of optimal decoding on choice correlations

Next we consider what happens when behavior is optimal given the neural response properties. For a fine discrimination task, the optimal decoder is $\mathbf{w} \propto \Sigma^{-1}\mathbf{f}'$ (Salinas and Abbott 1994, Moreno-Bote, Beck et al. 2014), where Σ is the true covariance matrix of neural population responses. For this decoder, the results of (Haefner, Gerwinn et al. 2013) imply that choice correlation for neuron k should be proportional to that neuron's sensitivity, $1/\theta_k$. This is not sufficient to demonstrate optimality, however: such a pattern of choice correlations indicates optimal decoding if and only if the proportionality constant equals the threshold of population activity, θ , which is the same as the behavioral threshold when decoding is optimal. All other values of the proportionality constant reflect suboptimal decoding, as we show later. For optimal computation, therefore, the choice correlations take the remarkably simple form of a ratio of discrimination thresholds (Methods):

$$C_k^{\text{opt}} = \frac{\theta}{\theta_k} \quad (3)$$

A consequence of this relationship is that more informative neurons — those with lower thresholds — have a greater correlation with behavior, as often observed experimentally (e.g. (Britten, Newsome et al. 1996, Purushothaman and Bradley 2005, Gu, Angelaki et al. 2008)).

From Eq. 3, we see that choice correlations predicted from optimal decoding depend on neural thresholds but not on the preferred stimulus, whereas predictions from suboptimal decoding (Eq. 2) depend on the preferred stimulus and not explicitly on the neural threshold. Yet, even in Eq. 3, the choice correlations depends on the preferred stimulus because neural tuning curves tend to have their steepest slope, and thus lowest threshold θ_k , for stimulus values offset from the preferred stimulus. As a consequence, we might worry that the predictions for optimal decoding (Eq. 3) could be satisfied even for a population of neurons decoded suboptimally, leading us to mischaracterize a suboptimal system as optimal. Indeed, that nearly happens for a homogeneous population of neurons that is decoded by the factorial decoder (Figure 2A–C), for which the patterns of choice correlation predicted from optimal decoding are roughly similar to those observed for a suboptimal correlation-blind decoder. However, when tuning curve amplitudes are sufficiently heterogeneous, stimulus preferences are less correlated with neural thresholds, so the relationship between choice correlation and threshold is weaker (Figure 2D–F). And when neural properties are modeled to match experimentally measured diversity in multiple tuning parameters, the relationship between choice correlation and neural threshold is extremely weak for the suboptimal decoder (Figure 2G–I).

We verified that Eqs. 2 and 3 can be used to successfully identify the decoding strategy by simulating heterogeneous populations and a decoding strategy that is chosen to be either optimal or suboptimal. Figure 3 plots the simulated choice correlations, first against the relevant neural properties (Figure 3A) and second against the choice correlations predicted from those properties (Figure 3B). The choice correlations for optimal decoding are perfectly fit by the prediction for optimal decoding (Eq. 3), and not by the prediction for suboptimal decoding (Eq. 2); the reverse holds approximately for choice correlations generated by suboptimal decoding. The bottom-left panel of Fig. 3B also provides a direct comparison of the two predictions, since the horizontal axis indicates predicted choice correlations for the suboptimal decoder while the vertical axis gives choice correlations for the optimal decoder (which is the true decoder for that simulation). The predictions are only weakly correlated ($r=.26$, $p<.01$ Pearson correlation test), and quite clearly distinguishable.

Consequences of information-limiting noise

Interestingly, noise correlations do not appear explicitly in Eq. 3 because the optimal decoder has removed them to the extent possible. Nonetheless, their structure has enormous importance for both behavior and choice correlations because they determine the population threshold, θ . Quantitative models of population codes have typically measured this by computing the Fisher information J , whose inverse provides a lower bound on the variance of an unbiased estimator. Since we defined the discrimination threshold θ as the standard deviation of our estimator, therefore $\theta \propto J^{-1/2}$, and consequently higher information permits a lower discrimination threshold.

According to current models of population codes (Shamir and Sompolinsky 2006, Ecker, Berens et al. 2011), when noise correlations are greater for neurons with similar tuning curves (as often observed empirically, see Cohen and Kohn 2011), the amount of Fisher information contained in a population is extensive, growing with the number of neurons in the population (Figure 4, solid cyan line). This is because when tuning curves are heterogeneous, the noise has a different structure from the signal, and thus noise and signal can be distinguished. For a population of 1000 neurons, optimal decoding could produce a behavioral threshold roughly 30-fold (1000) smaller than the threshold of a typical neuron that is tuned to the stimulus, and this is much smaller than generally observed (Cohen and Newsome 2009). According to Eq. 3, the choice correlations would then be correspondingly tiny, and most likely not significantly different from zero given typical measurement uncertainty. This argument rules out models in which the brain has extensive information and decodes it optimally.

However, models involving diverse neural populations might radically overestimate the amount of information in a population because they do not account for how the sensory periphery limits the information provided to a large cortical network. If information is not available at the input, then it cannot be created later by adding more neurons. Those extra neurons may appear to add more signal strength, but they also inherit noise with the same structure as the signal. As a consequence, that noise is correlated in a very particular way. For the fine discrimination task that we examined, the signal is encoded in the change in the mean rate, \mathbf{f}' , so the relevant information-limiting noise covariance is proportional to $\mathbf{f}'\mathbf{f}'^T$, which we have described elsewhere as ‘differential correlations’ (Moreno-Bote, Beck et al. 2014). More generally, for any discrimination task, the difference in mean neural activity between two stimuli can be denoted \mathbf{f} , and the relevant information-limiting correlations are then proportional to $\mathbf{f}\mathbf{f}^T$. A single noise covariance matrix can also contain information-limiting correlations for many different tasks simultaneously (Moreno-Bote, Beck et al. 2014). Our theoretical treatment applies to these circumstances as well.

Regardless of their specific form, information-limiting noise correlations may hide essentially unnoticed amongst other noise correlations with higher variance that arise downstream (Moreno-Bote, Beck et al. 2014). These information-limiting correlations cannot be averaged away by appropriately pooling many neurons, because these noise fluctuations are indistinguishable from changes in neural activity induced by a real signal. For our fine discrimination task, the total noise then has a covariance that can be modeled as

$$\Sigma = \Sigma_0 + \varepsilon \mathbf{f}'\mathbf{f}'^T \quad (4)$$

where Σ_0 is a covariance matrix of noise that does not limit information and ε represents the variance of the information-limiting noise. These two components of the covariance have very different structures. The more easily observed Σ_0 component exhibits higher correlations for neurons with similar preferred stimuli, as observed in many experiments (Cohen and Kohn 2011, Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). This Σ_0 component appears as a banded structure in the covariance, with fine striations due to tuning

curve heterogeneity (Figure 1E back, Methods, Eqs. 8–10). In contrast, the information-limiting component has a blockier structure (Figure 1E front).

Regardless of how the population is decoded, information-limiting noise prevents the variance of any unbiased linear estimator from falling below ε . The sum of noise covariances in Eq. 4 manifests as a sum of noise variances in the decoded estimate:

$\sigma_s^2 = \sigma_{0s}^2 + \varepsilon$, where σ_{0s}^2 is the variance that would have been obtained without information-limiting noise (Methods, Eq. 12). Since $\sigma_{0s}^2 \geq 0$, we see that $\sigma_s^2 \geq \varepsilon$. This same relationship can be expressed in terms of (linear) Fisher informations, $J = 1/\sigma_s^2$ and $J_0 = 1/\sigma_{0s}^2$, yielding

$$J = \frac{1}{1/J_0 + \varepsilon} \quad (5)$$

Thus information-limiting noise prevents the decoded information from exceeding $1/\varepsilon$ (Figure 4, solid red line).

While no decoder can exceed this information limit, it is of course possible to do worse. The quality of a decoder is determined by how efficiently it eliminates the noise that is not information-limiting. If ε is small relative to σ_{0s}^2 , then performance will be greatly enhanced by learning decoding weights that eliminate as much noise as possible. However, if $\varepsilon \gg \sigma_{0s}^2$, there is relatively little to be gained by fine tuning the decoding weights. This is why, in the absence of information-limiting correlations ($\varepsilon = 0$), a suboptimal correlation-blind decoder loses the vast majority of available information (Figure 4, dashed cyan curve; Supplemental Material), yet in the presence of information-limiting noise it loses only a modest fraction of the information that is available (Figure 4, dashed red curve). This demonstrates that large population codes with limited information are redundant and exhibit considerable robustness to suboptimal decoding, meaning that a broad range of decoders may all produce similar near-optimal performance.

Despite the importance of information-limiting correlations, they are difficult to estimate directly, requiring large simultaneous recordings with many trials. There are two reasons for this. First, the information-limiting component can be very small yet have enormous effects on population information. Second, the fine details of the correlation patterns matter greatly. Extrapolating the full noise correlations from a sparse subset of pairwise correlation measurements is extremely difficult, and mistakes can radically change the estimated information content of a neural population (Moreno-Bote, Beck et al. 2014). Fortunately, we show below that there are indirect consequences of this information-limiting noise that are observable with only single-neuron measurements: choice correlations should be observably large and should obey the predictions of optimal decoding (Eq. 3).

Choice correlations are influenced by both suboptimal decoding and information-limiting noise correlations according to the weighted sum

$$C_k \approx \alpha C_k^{\text{opt}} + \sqrt{1 - \alpha} C_k^{\text{sub}} \quad (6)$$

where $\alpha = \varepsilon J$ is the fraction of the uncertainty in the stimulus \hat{s} caused by the information-limiting noise (Methods). We emphasize that the decoder of Eq. 6 is not somehow both optimal and suboptimal. Instead, it is suboptimal for all $\alpha < 1$, but its choice correlations are a weighted sum of C_k^{opt} and C_k^{sub} , the choice correlations from Eqs. 3 and 2 for purely noise-limited optimal decoding or purely suboptimal decoding (of any type, not just correlation-blind) of a population with extensive information ($\varepsilon = 0$), respectively. As long as the behavioral threshold is primarily limited by noise and not by losses from suboptimal decoding, then α will be near 1. There will then be an inverse relationship between a neuron's threshold and its influence on behavior, regardless of neural correlations or the form of the decoder weights.

Suboptimal decoders can produce choice correlations that are scaled versions of optimal choice correlations

Surprisingly, one can identify circumstances in which the choice correlations have the same pattern as for optimal decoding, that is $C_k = \beta C_k^{\text{opt}}$, but with $\beta > 1$. This arises when behavior is driven by another, more informative, source of sensory signals beside the neural population under study. If responses of this more sensitive population are correlated with the observed population, then the observed neurons will exhibit choice correlations. As detailed in the Methods (Eqs. 24–29), $\beta > 1$ occurs only for suboptimal decoding, and can be explained quantitatively if the observed population is mostly ignored while the other more sensitive population is decoded efficiently.

Intuition for this result can be gained by examining results for a pair of idealized neurons, x and y , and then generalizing to two large populations. Imagine that the behavior is determined solely by the activity of neuron x , given by $x = s + n$ for signal s and noise n . Naturally, this decoded neuron will then be perfectly correlated with behavior. Since the neural threshold θ_x is the same as the behavioral threshold θ , the choice correlation is accurately described by $C_x = 1 = \theta / \theta_x$. Now imagine that neuron y carries the same stimulus-related signal and the same noise on every trial, except that the noise is multiplied by a factor of 2, $y = s + 2n$. Even though y is not decoded, it is still perfectly correlated with behavior since it is perfectly correlated with x . Yet because its neural threshold is twice as large due to greater noise, $\theta_y = 2\theta_x$, its choice correlation will be twice as large as an optimally decoded neuron with the same threshold: $C_y = 1 = 2(\theta / \theta_y)$.

The situation for two larger populations is analogous, albeit with the relevant signal and noise being distributed among many neurons along the direction of \mathbf{f}' . When population x is decoded near-optimally, its noise is correlated with behavior according to (Eq. 3). A second population y can have partially overlapping information, for instance inherited from overlapping subsets of upstream neurons (Figure 5A). This corresponds to partially correlated information-limiting noise. Figure 5B shows how this noise corresponds to trial-to-trial shifts in the hills of activity, with different variances in each population and a nontrivial covariance between them. When this covarying information-limiting noise is a larger fraction of the signal in population y compared to population x (i.e. its noise is larger in units of \mathbf{f}'), then choice correlations for neurons in population y will have a larger overall scale, but the same pattern, as choice correlations for neurons in population x (Figure 5C).

An important observation about this scenario is that choice correlations proportional but not equal to the optimal prediction (Eq. 3) occur only with some degree of suboptimal decoding. This can be seen in Figure 5D, which shows the efficiency (color scale) of a decoder that ignores the less informative of the two populations, where efficiency is the ratio of information actually decoded to the information that could have been decoded. This efficiency depends on the relative information content in the two populations (horizontal axis) as well as on β (vertical axis), which specifies how much choice correlations are amplified in the undecoded population relative to the prediction from optimal decoding (Eq. 3). Whenever the efficiency is 1 (white area of Figure 5D), then $\beta=1$ and the choice correlations match the optimal predictions. Conversely, whenever the choice correlations have $\beta \neq 1$, the efficiency is less than 1. The counter-intuitive situation in which choice correlations are greater than expected ($\beta > 1$) likely arises when there are two correlated populations of neurons that carry task-related information and the population with greater noise variance is decoded suboptimally (e.g., ignored).

Inferring decoding quality from neural data

To determine whether the information in a neural population is limited by noise or by suboptimal decoding, we can use linear regression to fit the measured choice correlations against those predicted from optimal and suboptimal decoders, while carefully accounting for uncertainties in each measurement (Methods, (Minka 1999)). We consider the two natural forms of suboptimality described above: correlation-blind decoding, and a decoder that ignores one population. These decoders generate choice correlations of the form

$C_k = \beta C_k^{\text{opt}} + \gamma C_k^{\text{cb}}$ (Eqs. 2, 6, 28–29). We fit these coefficients separately to allow for both of these forms of suboptimality. The coefficient β should reveal the fraction of behavioral variance caused by information-limiting noise in the recorded population.

To validate this approach, we simulated virtual neural populations and their virtual behavioral outputs, and used our method to try to recover the true decoder properties under realistic experimental conditions. These four model systems were: (i) optimal decoding and information-limiting noise, (ii) suboptimal, correlation-blind decoding of a population with extensive information, (iii) suboptimal, correlation-blind decoding with information-limiting noise, and (iv) two sub-populations with correlated information-limiting noise where only the more informative sub-population is decoded while we recorded from the other sub-population (Methods). For (iii), we set parameters such that $\alpha = 0.9$ in Eq. 6, meaning that 90% of the variance of the stimulus estimate was due to information limiting correlations. We then simulated recordings from small subsets of the virtual neurons, including measurement error, and estimated choice correlations, neural thresholds, tuning curves, and their corresponding uncertainties. The thresholds and tuning data were used to predict choice correlations according to Eqs. (2) and (3) separately, and these predictions were combined through linear regression to find the coefficients, β and γ , attached to the optimal and suboptimal choice correlations. Figure 6 shows these coefficients plotted separately (Figure 6A) and together (Figure 6B). In all four model systems, this method was able to recover the true values of both coefficients within the simulated experimental uncertainty. Figure 6C shows scatter plots of the predicted versus measured choice correlations, for three different predictors: optimal decoding (the psychophysical/neural threshold ratio, Eq. 3), suboptimal

correlation-blind decoding (Eq. 2), and the linear combination of the two that uses best-fit weights β and γ . As expected, the best predictors are the ones that match the actual decoding model (Figure 6C). We assessed model quality for each predictor using the Akaike Information Criterion (Methods), which measures the model log-likelihood while penalizing free parameters. We expect that the highest quality model will match the simulation's assumptions: the optimally-decoded simulation should be best fit by the optimal predictor, the correlation-blind simulation should be best fit by the correlation-blind predictor, etc. Indeed this is what we observe in Figure 6D. This demonstrates that our procedure successfully identifies whether neural activity is decoded optimally or suboptimally and can recover the fraction of behavioral variance caused by information-limiting noise.

Neural response properties during vestibular heading discrimination are consistent with near-optimal decoding

We now apply these theoretical insights to experimental data. In particular, we examine response properties and choice correlations of neurons recorded in multiple cortical and sub-cortical brain areas during a vestibular heading discrimination task (Gu, DeAngelis et al. 2007, Gu, Angelaki et al. 2008). Monkeys were translated forward and slightly leftward or rightward on a motorized platform, and were trained to report their perceived heading, left or right relative to straight forward, by making an eye movement to one of two targets. During this task, single neurons were recorded from the vestibular and cerebellar nuclei (VN/CN), the dorsal medial superior temporal (MSTd) area, and the ventral intraparietal (VIP) area. For each brain area, neural responses and behavioral choices were analyzed to extract choice correlations, neural thresholds and behavioral thresholds (Methods). In a separate stimulus condition, monkeys were translated along 8 equally-spaced headings in the horizontal plane during a visual fixation task, and neural responses were used to generate heading tuning curves and extract heading preferences. Measurements of pairwise noise correlations were also extracted from these data, as described previously (Gu, Liu et al. 2011, Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). Together, these data were used to test the theoretical predictions derived above.

Again we used linear regression to fit coefficients for optimal and suboptimal predictors of choice correlation to determine whether choice correlations were better predicted by optimal or suboptimal decoding. For VN/CN data obtained from 4 animals, the resultant regression coefficients on the suboptimal predictors were near zero, whereas those on optimal predictors were close to one (Figure 7A,B). We conclude that the information available in the vestibular and cerebellar nuclei is used near-optimally, and that behavioral performance is limited primarily by correlated noise, not suboptimal decoding. Likewise, weights for area MSTd in one monkey were consistent with optimal decoding, whereas choice correlations for MSTd in the other monkey were inconclusive.

Data from area VIP in two animals revealed near zero coefficients for the suboptimal predictor, but interestingly showed coefficients on the optimal predictor that were substantially greater than one (Figure 7A,B). As described above, this suggests that the information available in VIP is insufficient to account for behavioral performance. Instead, this finding is consistent with a model in which VIP is not decoded for heading

discrimination but nonetheless contains information that is correlated with another area that is decoded near-optimally (see Figure 5). This account is also consistent with recent preliminary results showing that reversible chemical inactivation of VIP does not impair heading discrimination (Klier, Liu et al. 2013), despite the fact that VIP neurons exhibit large choice correlations (Chen, Deangelis et al. 2013). Similarly, another recent study of the parietal cortex (area LIP) also reports high choice correlations, yet inactivating this region again has no discernible effect on visual motion discrimination (Yates, Katz et al. 2014).

As expected from optimal decoding, we found that the ratio of psychophysical to neural thresholds (Eq. 3) did a reasonable job of predicting the measured choice correlations (Figure 7C, top panel) for neurons in VN/CN and MSTd. We emphasize that this prediction has no free parameters. In contrast, the prediction of suboptimal decoding (Eq. 2) was poorly matched to data across all areas (Figure 7C, middle panel). The best linear combination of predictors using the regression weights above could provide a better explanation than either predictor alone, but the improvement over the parameter-free optimal decoding predictor was small for VN/CN in all animals and MSTd in one monkey (Figure 7C, bottom panel). To measure model quality, we computed the Akaike Information Criterion for the optimal model, the suboptimal (correlation-blind) model, and the best linear combination of those two models. For completeness we also test two additional models, one that fits the best scaling of the optimal predictor according to Eq. 28, and a null model that attributes all variation in measured C_k to random chance (Figure 7D). According to this statistical measure, our data was strong enough to significantly differentiate between the models.

We also examined three additional neural response properties for possible deviations from optimality. First, for optimal linear computation, choice correlations should be zero for neurons with very high thresholds. A few of these uninformative neurons did have choice correlations that differed from zero by more than two standard deviations of measurement uncertainty, but no more than expected by chance ($p=.10$, t -test on C_k/σ_{C_k} for the 19 neurons with unmeasurably large thresholds). Second, on the other end of the sensitivity spectrum, optimal computation requires that no neuron has a better threshold than the behavior. Indeed, our data, like those of a previous study (Cohen and Newsome 2009), reveal that no neurons have a threshold lower than the animal's behavioral threshold, as long as neural thresholds are properly corrected for use of a neuron-antineuron pair (a correction that was not applied in (Cohen and Newsome 2009)). Third, there should be no substantially negative choice correlations; if a neuron prefers one stimulus polarity (leftward or rightward heading), it should not drive behavior toward the opposite choice. Although we did observe some negative choice correlations, the associated 95% confidence intervals exclude zero from below for only 4/339 recorded cells, a proportion which is not significant ($p=.96$). All of these lines of evidence are consistent with the idea that, on this simple heading discrimination task, the brain uses the vestibular information in its neural populations near-optimally.

To summarize, our results indicate that areas MSTd and VN/CN provide redundant codes that are read out near-optimally whereas our data suggest that information in area VIP is not used efficiently for this task. From the pattern and scale of choice correlations in VIP, we

infer that VIP must be highly redundant with VN/CN, MSTd, or another unrecorded area, in which case the brain loses very little information by ignoring VIP.

Discussion

The high information content in current models of neural population codes (Shamir and Sompolinsky 2006, Ecker, Berens et al. 2011) would lead to a huge difference between neural thresholds and behavioral thresholds and immeasurably tiny choice correlations if the brain used all of that information efficiently. Since choice correlations are often not small, there are two alternatives: either the models are incorrect with regard to the high information content of population codes, or the brain is highly suboptimal in extracting the information. Our analysis revealed that these two alternatives have distinguishable consequences, and we tested for these consequences in different brain regions involved in vestibular judgments of self-motion. Our results show that, at least for the simple discrimination task considered here, the first alternative is more likely: the brain has limited information and is able to extract it near-optimally. This is a crucial property of neural coding that must be considered in future theories and experiments.

Achieving near-optimal performance may not be such an impressive feat, as our results imply that the information encoded in neural populations is highly redundant and therefore robust to suboptimal decoding. The information-limiting noise correlations cannot be removed because they look just like the signal. In their presence, there is little advantage to optimally removing all of the remaining components of correlated noise. As a result, a broad range of suboptimal decoding weights are about equally good as long as they produce less variance than the information-limiting correlations. Optimality measures the quality of decoding, rather than the particular network structure used to decode. Thus the true decoding weights could resemble those of a correlation-blind decoder, or many other decoders, and yet be near-optimal due to the effect of the information-limiting correlations.

In an early study of correlations between neural activity and behavior (Zohary, Shadlen et al. 1994), the authors wrote that “The covariation of single-neuron responses and psychophysical decisions, an observation that strains credulity at first glance, is a logical consequence of weakly correlated noise within the pool of sensory neurons leading to the decision.” Our study shows that this insight remains essentially true—correlated noise can create significant choice correlations. However, weak positive noise correlations do not always produce large choice correlations; this is only true for the model they consider with homogeneous neurons and homogeneous noise correlations. In general, the only noise correlations that lead to significant choice correlations are those that mimic the effect of the stimulus on population activity. For fine discrimination tasks, we have called these differential correlations (Moreno-Bote, Beck et al. 2014) because the correlations mimic the relevant signal which is the differential (derivative) of the mean responses.

Suboptimal decoding, and more generally suboptimal neural computations, can also generate large choice correlations. However, the present results support the conclusion that signals from the VN/CN and MSTd are decoded near-optimally. VIP exhibits high choice probabilities and robust noise correlations much like VN/CN, yet our analysis suggests that

VIP is decoded suboptimally, if at all. However, the consequences of this suboptimality for the animal are modest because, according to our results, VIP contains little information that is not already available in the population responses of VN/CN and MSTd or other areas.

A recent study (Haefner, Gerwinn et al. 2013) demonstrated how to reconstruct the weights that the brain gives to different neurons from choice probability measurements and correlated noise. Despite the elegance of this result, one consequence of our findings is that such a full reconstruction may not be feasible in the presence of information-limiting correlations. Due to the robustness of the resultant neural code, many different decoding weight profiles can extract nearly all of the available information (Fig. 4). Under conditions in which behavior is limited by noise (α near 1), Eq. 6 shows that choice correlations will be dominated by the optimal term αC_k^{opt} regardless of the particular structure of the decoding weights, while the suboptimal term $\sqrt{1 - \alpha} C_k^{\text{sub}}$ will be small. Any resultant deviations of choice correlations from the optimal predictor are therefore likely to be subtle and difficult to detect in the presence of measurement noise. Since many weight profiles would generate almost the same choice probabilities, it may not be possible to use choice probabilities to compute decoding weights from experimental data using the approach of Haefner et al. (2013).

Despite these challenges in extracting decoding weights from neural and behavioral data, our results show that one can still fruitfully compare the patterns of choice correlations expected under different hypotheses. By analyzing choice correlations generated by optimal linear decoding, correlation-blind decoding, or a decoder using only a subset of all neurons, we are able to draw strong conclusions about information processing in the brain. Due to measurement noise in the data, it remains possible that neural processing is even better described by some other class of suboptimal decoders that we did not consider. However, we presented theoretical arguments in favor of limited information. First, if cortical populations have extensive information then the brain would need to throw away almost all of it to explain behavioral performance. Second, since cortical populations are much larger than the sensory population but not much noisier, the extensive information model attributes more information to the cortex than to the sensors — and this is prohibited by the data processing inequality. In addition to these theoretical arguments, the predictions of the limited-information model provide a good match to data. Thus we conclude that the vestibular code for heading is redundant, that the information in VN/CN is used near-optimally, and that although VIP contains robust vestibular information, it is likely not decoded efficiently for this heading discrimination task. This would predict that deactivating VIP should have little effect on the performance of the animal — which is indeed what we have recently found experimentally (Lakshminarasimhan, Liu et al. 2014).

Our analytical results reveal several other properties of choice correlation that defy conventional wisdom. One common expectation is that choice correlation should be smaller for larger populations since information is distributed across more neurons. However, if the code is redundant and decoded near-optimally, then the choice correlations do not vary with population size. This is because information saturates with the number of neurons, so the

ratio of the population threshold to an individual neuron's threshold saturates too. According to Eq. 3, the choice correlations are equal to this ratio, and therefore they saturate as well.

Another expectation based on previous results is that choice correlations should be greater when noise correlations are larger (Nienborg and Cumming 2010, Liu, Gu et al. 2013). This is true only for a specific pattern of correlations, or for certain types of decoders. For suboptimal decoding of highly informative populations, choice correlations do increase with the overall scale of correlated noise (Eq. 2). However, our results instead favor near-optimal decoding, for which most correlations play little role: only information-limiting correlations increase the choice correlations according to Eq. 3. Since these special noise correlations may constitute only a small portion of the total measured noise covariance, choice correlations could potentially rise even while the total noise correlation falls.

A consequence of optimal decoding that emerges from Eq. 3 is that choice correlations should depend on behavioral performance: neurons with similar stimulus sensitivities should have higher choice correlations in animals that have higher behavioral thresholds. Conversely, if different brain regions seem to have different choice correlations, one must take care to account for differences in psychophysical sensitivity between animals before concluding that the difference between brain areas is meaningful. The surest diagnostic of differences between brain regions would be to measure choice correlations from multiple areas within a single animal.

In this study, we evaluated whether choice correlations are consistent with optimal or suboptimal decoding. These results depend on the particular class of suboptimal decoders we considered in the extensive information case, namely the correlation-blind decoders. This class is both biologically plausible, well-established in the literature, and quite general, although not all-encompassing. As long as neurons are broadly tuned, noise correlations resemble signal correlations, and the suboptimal decoder does not remove these broad correlations, then the resultant pattern of choice correlations will be close to a sinusoidal function of the preferred direction of the neurons (Eq. 2, Supplemental Fig. S3). Nonetheless, it is possible for a suboptimal decoder to produce patterns of choice correlations that differ substantially from Eq. 2. For instance, a decoder might successfully remove the noisiest and lowest-frequency noise mode, but fail to suppress the next-noisiest mode, which would generate both positive and negative choice correlations in a pattern approximately proportional to $\text{sgn}s_k \sin 2s_k$. While it would be surprising if the brain were to successfully learn to avoid decoding the noisiest mode but not other highly noisy modes, we cannot definitively rule it out. On the contrary, we can always concoct some combination of correlations and suboptimal decoders that would be consistent with fine details of our measurements. However, this would require unpalatable fine-tuning of the model, and thus would be rejected in favor of the simple limited-information model we offer here, which fits the data quite well.

Our results also depend on models of neural signals and neural noise. Our predictions of choice correlations under suboptimal decoding rely on the structure of noise correlations. We modeled the dominant noise correlations as proportional to signal correlations, a trend that is generally supported by data (Liu, Gu et al. 2013). Note that task-dependent changes

in correlation amplitude (Cohen and Newsome 2008) or increases in differential correlations (Bondy and Cumming 2013) will not change the overall shape of choice correlations, and thus will not qualitatively alter the fact that observed choice correlations are inconsistent with decoders that ignore correlations.

On the other hand, the predictions of optimal decoding do depend on how we measure neural and behavioral thresholds. Ideally, we would like to compute neural thresholds from the same neural signals that drive the behavior, but we don't know precisely what those signals are. For example, some past analyses comparing neurons and behavior (Britten, Shadlen et al. 1992) integrated neural signals over time windows that likely extend past the animal's decision time, artificially inflating neural sensitivity relative to behavior (Cohen and Newsome 2009). Our analysis assumed that the relevant neural signals are the spike counts in a 400 msec time window around the peak movement velocity, rather than the full 2sec movement duration or the central 1sec where the gaussian movement profile is most substantial. However, many neurons in VN/CN and VIP have some sensitivity to acceleration which is small at that time (Chen, DeAngelis et al. 2011, Liu, Gu et al. 2013), so their threshold may be lower slightly before our chosen time window. Consequently, we tested how sensitive our results would be to a global decrease in neural threshold by 2, such as might be caused by advancing or doubling the temporal integration window. The result was an equal decrease in the weight given to the optimal decoding predictor without a concomitant change in the suboptimal predictor weight. Since most inferred optimal weights were slightly greater than 1, this change brings the weights slightly below but still within the 95% confidence intervals around 1. Thus, our conclusions are robust to moderate errors in estimating neural sensitivity. It is also possible that the brain might use a more complex temporal weighting function of the neural response than just a simple sum (Drugowitsch, DeAngelis et al. 2014). It will be interesting to explore such temporal aspects of neural processing using generalizations of choice measures like psychophysical or neurometric kernels (Nienborg and Cumming 2009, Wohrer and Machens 2013), especially in a more natural setting where the relevant signal itself changes over time.

In summary, we have presented a theory of how choice correlations depend on the information content of a neural population. A large cortical population that has access to only limited information from its sensors will exhibit a specific form of noise correlations that are difficult to detect yet have an enormous impact on the neural code. One consequence of these information-limiting noise correlations is that a large class of decoders can then extract information near-optimally. Many neurons will then be correlated with behavior because they are correlated with each other. Our data provide evidence that this is the situation in the vestibular system during heading discrimination tasks. These theoretical and experimental conclusions highlight the importance of understanding the detailed structure of noise correlations, for these details fundamentally change how the brain can use and process sensory information.

Methods

Subjects and apparatus

Eight rhesus monkeys (*Macaca mulatta*, 4–6 kg) were chronically implanted with an eye coil, head-restraint ring, and a plastic grid of holes through which guide tubes were passed for electrophysiological recordings (Meng, Green et al. 2005, Gu, Watkins et al. 2006). All surgical and experimental procedures were approved by the Institutional Animal Care and Use Committee at Washington University and were performed in accordance with institutional and NIH guidelines. Motion stimuli were delivered using a six-degree-of-freedom motion platform (Moog 6DOF2000E), as described previously (Gu, Watkins et al. 2006).

Vestibular heading discrimination task

Animals were trained to perform a fine heading discrimination task around psychophysical threshold. During neural recordings in the discrimination task, seven logarithmically spaced headings ($\pm 6.4^\circ$, $\pm 2.6^\circ$, $\pm 1^\circ$, 0° relative to straight ahead) were presented in a block of randomly-interleaved trials, while animals maintained fixation on a head-fixed target at the center of the display ($2^\circ \times 2^\circ$ electronic window). The range and spacing of headings were chosen carefully to obtain near-maximal psychophysical sensitivity while allowing neural sensitivity to be reliably estimated for most neurons. The motion trajectory (30 cm displacement) was 2 sec in duration and followed a Gaussian velocity profile (SD: 0.5 sec; peak velocity: 45 cm/sec), with a corresponding biphasic linear acceleration profile ($\pm 0.1G = \pm 0.98 \text{ msec}^{-2}$).

At the end of each trial of the discrimination task, the fixation point disappeared, two choice targets appeared, and the monkey was trained to make a saccade to the left or right target to report his perceived heading (leftward or rightward relative to an internal standard of straight ahead). Correct choices were rewarded with a drop of water or juice. For the ambiguous straight-forward heading direction (0°), rewards were delivered randomly on half of the trials. If fixation was broken at any time during the 2 sec motion stimulus, the trial was aborted and the data were discarded. If neural isolation was lost before completion of at least 10 repetitions of the discrimination task, that neuron was excluded from quantitative analysis. In our sample, cells were held long enough to be tested with at least 20 repetitions of each distinct stimulus for 77% (75/97) of neurons in VN/CN, 95% (174/183) of neurons in MSTd, and 69% (41/59) of neurons in VIP.

Neural recordings

We recorded extracellularly the activity of single neurons in the VN/CN, MSTd, and VIP using epoxy-coated tungsten microelectrodes (FHC, 5–7 $M\Omega$ impedance for VN/CN, 1–2 $M\Omega$ for MSTd and VIP). To target recordings to the VN and CN, we first identified the abducens nuclei bilaterally in initial experiments with each animal. We then used the location of the abducens nuclei to guide electrode penetrations into the CN and VN (Meng, Green et al. 2005, Liu, Gu et al. 2013). Area MSTd was located using a combination of magnetic resonance imaging (MRI) scans, stereotaxic coordinates (~ 15 mm lateral and ~ 3 –6 mm posterior to AP-0), white/gray matter transitions, and physiological response

properties. In some penetrations, electrodes were further advanced into the retinotopically organized area MT. Most recordings concentrated on the posterior/medial portions of MSTd, corresponding to more eccentric, lower hemifield receptive fields in the underlying area MT (Gu, Watkins et al. 2006, Gu, DeAngelis et al. 2007). To localize area VIP, we first identified the medial tip of the intraparietal sulcus and then moved laterally until there was no longer directionally selective visual response in the multiunit activity. At the anterior end, visually responsive neurons gave way to purely somatosensory cells in the fundus. At the posterior end, direction-selective neurons gave way to visual cells that were not selective for motion (Chen, DeAngelis et al. 2011).

Computing choice correlations and thresholds

Behavioral performance was quantified by plotting the proportion of ‘rightward’ choices as a function of heading (the azimuth angle of translation relative to straight ahead). Psychometric data were fit with a cumulative Gaussian function,

$P(\text{right}|s) = \frac{1}{2} \text{erfc} \left(\frac{(\mu - s)/\sigma}{\sqrt{2}} \right)$ where $P(\text{right}|s)$ is the proportion of choices of rightward headings, s is the actual stimulus direction, μ is the mean of the Gaussian (corresponding to the point of subjective equality) and σ is the standard deviation. Psychophysical threshold was defined as the standard deviation of the Gaussian fit, σ , which corresponds to 68% correct performance (assuming no bias).

For the analyses of neural responses, we used mean firing rates calculated during the middle 400ms interval of each stimulus presentation. To characterize neural sensitivity, we used ROC analysis to compute the ability of an ideal observer to discriminate between two oppositely-directed headings (e.g., -6.4° vs. $+6.4^\circ$) based solely on the firing rate of the recorded neuron and a presumed ‘antineuron’ with opposite tuning (Britten, Shadlen et al. 1992). ROC values were plotted as a function of heading, resulting in neurometric functions that were also fit with a cumulative Gaussian function. Neural threshold was defined as the standard deviation of the fitted Gaussian, but increased by a factor of $\sqrt{2}$ to account for the extra information from the antineuron. For a handful of insensitive neurons for which the estimated thresholds were very large, values were truncated to an upper limit of 300° . Because neural and psychophysical thresholds were measured simultaneously during each experimental session, the two could be directly and quantitatively compared.

To quantify the relationship between neural responses and the monkey's perceptual decisions, we also computed ‘choice probabilities’ using ROC analysis (Green and Swets 1966). For each heading, neural responses were sorted into two groups based on the choice that the animal made at the end of each trial: ‘preferred’ choices refer to decisions that favor the preferred heading of the recorded neuron, whereas ‘null’ choices refer to decisions in the opposite direction. ROC values were calculated from these response distributions, yielding a choice probability (CP) for each heading, as long as the monkey made at least 3 choices in favor of each direction. To combine across different headings, we computed a grand CP by balanced z -scoring of responses in different conditions, combining z -scored response distributions across conditions, and then performing ROC analysis on those distributions (Kang and Maunsell 2012). The statistical significance of CPs (i.e., whether they were significantly different from the chance level of $1/2$) was determined by a permutation test

(1000 permutations). Finally, experimental choice probabilities were converted into choice correlations according to $C_k = \frac{\pi}{\sqrt{2}}(CP_k - \frac{1}{2})$ ((Haefner, Gerwinn et al. 2013), Methods below).

Modeling neural responses and correlations

We characterize neural responses \mathbf{r} during discrimination tasks by their stimulus-dependent mean $\mathbf{f}(s)$, also known as the tuning curve, and the covariance of their noise given the stimulus, $\Sigma(s)$: $\mathbf{r} \sim N(\mathbf{f}(s), \Sigma(s))$.

The tuning curves are approximated by baseline-shifted von Mises functions,

$$f_k(s; \psi_k) = b_k + a_k \exp[\kappa_k (\cos(s - s_k) - 1)] \quad (7)$$

with parameters $\psi_k = \{b_k, a_k, \kappa_k, s_k\}$ representing baseline firing rate, modulation amplitude, width, and preferred heading respectively (Figure 1A).

We assume that noise correlation coefficients R , in the absence of information-limiting noise, are proportional to the signal correlation coefficients on average,

$$\bar{R} = (1 - c_0)I + c_0 R^{\text{sig}} \quad (8)$$

where R^{sig} is the correlation coefficient matrix between mean neural responses $\mathbf{f}(s)$ over a uniform distribution of inputs $p(s)$. This model is consistent with prior descriptions of measured neural correlations (Cohen and Kohn 2011, Chen, Deangelis et al. 2013, Liu, Gu et al. 2013). For the tuning curves given above by (7), the signal correlations are given by

$$\begin{aligned} R_{ij}^{\text{sig}} &= \frac{\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle}{\sqrt{(\langle f_i^2 \rangle - \langle f_i \rangle^2)(\langle f_j^2 \rangle - \langle f_j \rangle^2)}} \\ &= \frac{I_0(\sqrt{\kappa_i^2 + \kappa_j^2 + 2\kappa_i \kappa_j \cos(s_i - s_j)}) - I_0(\kappa_i)I_0(\kappa_j)}{\sqrt{(I_0(2\kappa_i) - I_0(\kappa_i)^2)(I_0(2\kappa_j) - I_0(\kappa_j)^2)}} \end{aligned} \quad (9)$$

where I_0 is a modified Bessel function of the first kind. Note that this expression depends on the difference between the preferred headings of a pair of neurons, $s_i - s_j$ (Figure 1B).

To introduce diversity into the resultant covariances, we first generated a mean covariance

matrix that scaled with firing rates according to $\left(\bar{\Sigma}_0\right)_{jk} = \sqrt{f_j f_k} \bar{R}_{jk}$. We then drew a covariance matrix from a Wishart distribution with mean $\bar{\Sigma}_0$ and $2N$ degrees of freedom, where N is the number of neurons:

$$\Sigma_0 \sim w(\bar{\Sigma}_0, 2N)/2N \quad (10)$$

The diversity in the covariance matrix Σ_0 induced corresponding diversity in the correlation coefficient matrix according to $R_{jk} = (\sum_0)_{jk} / \sqrt{f_j f_k}$ (Figure 1B), such that the correlation coefficient between R and R^{sig} was around 0.8 for $N=500$. Supplemental Figure S2B shows that generating even greater diversity in correlations does not change our main results.

For simulations with information-limiting correlations, we added a component to the covariance given by $\epsilon \mathbf{f} \mathbf{f}^T$ (Moreno-Bote, Beck et al. 2014), or given by the rank-two variant specified below (Eq. 26).

Linear decoding

We model decoding as unbiased linear estimation of the stimulus from the population activity, $\hat{s} = \mathbf{w}^T (\mathbf{r} - \mathbf{f}(s_0))$ where \mathbf{w} is a vector of the decoder weights and s_0 is the reference stimulus around which discriminations are made (Figure 1C,D). Binary behavioral choices are generated from this estimate according to $\text{sgn}(\hat{s} - s_0)$. Unbiased decoding requires that $\langle \hat{s} \rangle_{p(\mathbf{r}|s)} = s$, leading to the constraint that $\mathbf{w}^T \mathbf{f}' = 1$. The variance of this estimator is given by

$$\sigma_{\hat{s}}^2 = \mathbf{w}^T \Sigma \mathbf{w} \quad (11)$$

Optimal decoding under these assumptions is achieved by decoding weights given by $\mathbf{w} \propto \Sigma^{-1} \mathbf{f}'$ (Salinas and Abbott 1994, Moreno-Bote, Beck et al. 2014), where $\mathbf{f}' = d\mathbf{f}/ds$ and the proportionality is determined by the condition $\mathbf{w}^T \mathbf{f}' = 1$ required for unbiased decoding. This optimal decoder then gives a variance of

$\sigma_{\hat{s}}^2 = \mathbf{w}^T \Sigma \mathbf{w} = \mathbf{f}'^T \Sigma^{-1} \Sigma \Sigma^{-1} \mathbf{f}' / (\mathbf{f}'^T \Sigma^{-1} \mathbf{f}')^2 = 1 / \mathbf{f}'^T \Sigma^{-1} \mathbf{f}'$, which is precisely equal to the inverse of the linear Fisher information, $J = \mathbf{f}'^T \Sigma^{-1} \mathbf{f}'$.

More generally, we define the linear Fisher information of a (possibly suboptimal) estimator to be the inverse variance of that estimator, $J = 1 / \sigma_{\hat{s}}^2$. When decoding is optimal, the linear Fisher information of the decoder is equal to the linear Fisher information for the whole population.

To compute the variance of an unbiased estimator in the presence of information-limiting noise, we use the decomposition of Eq. 4 to obtain

$$\sigma_{\hat{s}}^2 = \mathbf{w}^T (\Sigma_0 + \epsilon \mathbf{f} \mathbf{f}^T) \mathbf{w} = \sigma_{0\hat{s}}^2 + \epsilon \quad (12)$$

where $\sigma_{0\hat{s}}^2 = \mathbf{w}^T \Sigma_0 \mathbf{w}$ is the estimator variance for $\epsilon=0$. Here we used the fact that $\mathbf{w}^T \mathbf{f}' = 1$ for unbiased decoding, so $\mathbf{w}^T (\epsilon \mathbf{f} \mathbf{f}^T) \mathbf{w} = \epsilon$. Using the inverse relationship between linear Fisher information and estimator variance, we obtain

$$J = \frac{1}{1/J_0 + \epsilon} \quad (13)$$

where $J_0 = 1/\sigma_{0\hat{s}}^2$ is the information that would have been obtained from that decoder if $\varepsilon=0$.

One can also consider estimators $\hat{s}(r_k)$ based only on the k th single neuron response r_k , which have variance

$$\sigma_{\hat{s}(r_k)}^2 = \frac{\sigma_{\hat{s}(r_k)}^2}{(f'_k)^2} \quad (14)$$

where $f'_k = df'_k/ds$ is the derivative of the tuning curve with respect to the stimulus and $\sigma_{r_k}^2$ is the variance of the neuron's response.

Choice correlation

Choice correlation C_k is the Pearson correlation coefficient between the neural response and the behavioral choice \hat{s} . Under the model assumptions, the choice correlation for neuron k is given by

$$C_k = \text{Corr}(\hat{s}, r_k) = \frac{\langle \hat{s} r_k \rangle - \langle \hat{s} \rangle \langle r_k \rangle}{\sqrt{(\langle \hat{s}^2 \rangle - \langle \hat{s} \rangle^2)(\langle r_k^2 \rangle - \langle r_k \rangle^2)}} = \frac{(\sum \mathbf{w})_k}{\sqrt{\sum_{kk} \mathbf{w}^\top \sum \mathbf{w}}} \quad (15)$$

Previous experimental work on this topic has generally quantified the relationship between neural responses and choices by the choice probability (Britten, Shadlen et al. 1992), which is based on ROC analysis (Green and Swets 1966). These two are similar measures; in fact previous work (Haefner, Gerwinn et al. 2013) showed that the choice probability CP_k for

neuron k is nearly linearly related to the quantity $(\sum \mathbf{w})_k / \sqrt{\sum_{kk} \mathbf{w}^\top \sum \mathbf{w}}$, which we recognize as the choice correlation C_k :

$$\text{CP}_k \approx \frac{1}{2} + \frac{\sqrt{2}}{\pi} C_k \quad (16)$$

We can also directly calculate the correlation between neural response and a binary choice $\text{sgn}(\hat{s})$ instead of a continuous choice \hat{s} , and find it is exactly proportionality to (Eq. 15),

$\text{Corr}(\text{sgn}(\hat{s}), r_k) = \text{Corr}(\hat{s}, r_k) \sqrt{2/\pi} \approx 0.8 \text{Corr}(\hat{s}, r_k)$. Given the greater conceptual simplicity of choice correlation (Eq. 15), and its near-equivalence to choice probability, we prefer to use it in our analyses.

Choice correlations for optimal decoding

A locally optimal linear decoding \hat{s}_{opt} of a neural population uses the weights $\mathbf{w}_{\text{opt}} \propto \Sigma^{-1} \mathbf{f}'$ (Salinas and Abbott 1994, Moreno-Bote, Beck et al. 2014). We specify the weights only up to a proportionality factor because any rescaling of the decoder weights has no effect on the choice correlation, as the effect on numerator and denominator cancel. Substituting these weights into Eq. (15), we find choice correlations given by

$$C_k = \frac{(\sum \Sigma^{-1} \mathbf{f}')_k}{\sqrt{\sum_{kk} \mathbf{f}'^\top \Sigma^{-1} \mathbf{f}'}} = \frac{f'_k}{\sigma_k} \sigma_s = \frac{\theta}{\theta_k} \quad (17)$$

where neural and behavioral thresholds θ_k and θ are defined here as the standard deviation of estimators based on a single neuron response (Eq. 14) and the full population response (Eq. 11). Other definitions of threshold use a particular performance criterion such as 75% correct, but this merely introduces a proportionality constant for each threshold that cancels in the ratio of Eq. (17).

Choice correlations for suboptimal decoding

The factorized decoder uses $\mathbf{w} \propto \mathbf{F}^{-1} \mathbf{f}'$ where \mathbf{F} is a diagonal matrix with the noise variances on the diagonal, which is $F_{kk} = f_k$ for neurons whose variance scales with their mean. Substitution into the expression for choice correlation (Eq. 15) yields

$$C_k^{\text{cb}} = \frac{(R\mathbf{q})_k}{\sqrt{\mathbf{q}^\top R \mathbf{q}}} \quad (18)$$

where $\mathbf{q} = \mathbf{F}^{-1/2} \mathbf{f}'$ is the signal-to-noise ratio for each neuron.

The numerator $(R\mathbf{q})_k$ is a weighted sum over the signal-to-noise ratios of many neurons. For broad correlations, this averages away the tuning curve diversity, leaving only the mean $(R\mathbf{q})_k$ where an overbar designates an average over the tuning parameters $\{b, a, \kappa\}$. Similarly, the quadratic form $\mathbf{q}^\top R \mathbf{q}$ is approximately $\bar{\mathbf{q}}^\top R \bar{\mathbf{q}} + \sigma_q^2 \text{Tr} R$, and the former term dominates since it scales as N^2 rather than $\text{Tr} R = N$. Moreover, if the heterogeneity in correlations arises from a Wishart distribution (Eq. 10) then variability in R averages away just like the diversity in \mathbf{q} , leaving us with

$$C_k^{\text{cb}} \approx \frac{(\bar{R}\bar{\mathbf{q}})_k}{\sqrt{\bar{\mathbf{q}}^\top \bar{R} \bar{\mathbf{q}}}} \quad (19)$$

(see Supplementary Materials for details). This average correlation matrix is circulant, so $R_{ij} = R_{i-j}$, and is well-approximated by $R_{ij} = c_0 \exp(\kappa(\cos s - 1)) + (1 - c_0)\delta_{ij}$ where $s = s_i - s_j$. Circulant matrices are diagonalized in the Fourier basis, with eigenvalues given by the frequency spectrum R_ν for frequency ν . This is given by modified Bessel functions of the

first kind I_ν : $\bar{R}_\nu = \mathcal{F}[\bar{R}_{i-j}] \sim c_0 I_\nu(\bar{\kappa}) + \frac{1 - c_0}{N}$. These functions fall off super-exponentially rapidly with frequency, at least as $R_\nu \sim (\kappa/2)^\nu / \nu!$ (Amos 1974), and are thus strongly dominated by the lowest frequencies, particularly when tuning curves are broad, with $\kappa \sim 1$. We assume that the decoder, being unbiased, weights left and right choices antisymmetrically, which implies that the decoder weighting profile is an odd function of the neural sensitivity f'_k ; therefore so is \mathbf{q} . This cancels all symmetric patterns in the numerator $R\mathbf{q}$ of Eq. 19. Combined with the rapid falloff in the frequency spectrum, the dominant antisymmetric noise pattern is $\sin s_k$, with the next pattern at least 4 times weaker. Only if

the decoder attenuates this dominant lowest frequency to a level much smaller than the other frequencies will the resultant correlations differ substantially from the sinusoidal prediction of Eq. 2. Correlation-blind decoders that weight neural responses based solely on their single-neuron properties cannot generically remove this lowest frequency. In the Supplementary Material we present detailed calculations based on the particular functional form of the tuning curves (Eq. 7).

The correlation-blind decoder generates choice correlations similar to the leading antisymmetric eigenvector of R . For correlations with considerable diversity around R , this eigenvector won't look like a perfect sinusoid as was the case previously, but like a sinusoid with some added scatter, as seen in the choice correlations of Supplemental Figure S3B.

Choice correlations for suboptimal decoding with information-limiting correlations

Suboptimally decoding neural responses that have information-limiting noise correlations (Eq. 4) leads to choice correlations that can be expressed as a sum of two terms:

$$C_k = \frac{(\sum \mathbf{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\sum_0 \mathbf{w} + \epsilon \mathbf{f}' \mathbf{f}'^T \mathbf{w})_k}{\sigma_k \sigma_{\hat{s}}} \quad (20)$$

For unbiased decoding, $\mathbf{w}^T \mathbf{f}' = 1$ (see Linear decoding above). Some manipulation gives

$$C_k = \frac{(\sum_0 \mathbf{w})_k}{\sigma_{0k} \sigma_{0\hat{s}}} \frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} \frac{\sigma_{0k}}{\sigma_k} + \frac{f'_k}{\sigma_k} \sigma_{\hat{s}} \frac{\epsilon}{\sigma_{\hat{s}}^2} \quad (21)$$

where $\sigma_{0k} = (\Sigma_0)_{kk} \approx \sigma_k$ for small information-limiting noise variance (which has a large effect on information despite the small variance), and where $\sigma_{0\hat{s}}$ is the standard deviation of the estimate produced by the same suboptimal decoder \mathbf{w} in the absence of information-limiting correlations, i.e. when the noise covariance is Σ_0 . Since $\sigma_{\hat{s}}^2 = \sigma_{0\hat{s}}^2 + \epsilon$, we find that

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \epsilon / \sigma_{\hat{s}}^2} = \sqrt{1 - \alpha} \quad (22)$$

with $\alpha = \epsilon / \sigma_{\hat{s}}^2$. Substituting these into (15) we find that the choice correlation is a weighted sum of the choice correlations for optimal and suboptimal decoding

$$C_k \approx C_k^{\text{sub}} \sqrt{1 - \alpha} + C_k^{\text{opt}} \alpha \quad (23)$$

where C_k^{sub} and C_k^{opt} are, respectively, the choice correlations for suboptimal decoding with $\epsilon=0$, and optimal decoding using $\mathbf{w} \propto \Sigma^{-1} \mathbf{f}'$ where the covariance included enough information-limiting noise (Eq. 4) to match the observed performance. The same derivation applies to any other binary discrimination task in which the stimuli change the mean firing rates by \mathbf{f} (e.g., a coarse discrimination task in noise). The relevant information-limiting

correlations will then be proportional to $\mathbf{f} \mathbf{f}^T$ and the choice correlations will again be a weighted sum satisfying Eq. 23.

Choice correlations for optimal decoding of only a subpopulation

We now show that if neurons are divided into two populations that receive limited but overlapping information, and only one population is decoded optimally while the other is ignored, then choice correlations in the latter population can be proportional to the optimal choice correlations (Eq. 3) with a proportionality constant greater than 1.

We denote the tuning curves of the two populations by \mathbf{x} and \mathbf{y} , so the full tuning curves are $\mathbf{f}=(\mathbf{x},\mathbf{y})$, with neural responses on a single trial given by $\mathbf{r}=(\mathbf{r}_x, \mathbf{r}_y)$. For fine discrimination tasks, information-limiting noise in each population separately takes the form of differential correlations (Moreno-Bote, Beck et al. 2014), $\varepsilon_{xx}\mathbf{x}'\mathbf{x}'^T$ and $\varepsilon_{yy}\mathbf{y}'\mathbf{y}'^T$. If these two populations share information, then there will also be correlations between brain regions of the form $\varepsilon_{xy}\mathbf{x}'\mathbf{y}'^T$ (depicted in Figure 5A). We can express the full noise covariance Σ as a rank-two perturbation UEU^T to a covariance Σ_0 that does not limit information, where

$$U^T = \begin{pmatrix} -\mathbf{x}' & -0 \\ -0 & -\mathbf{y}' \end{pmatrix} \quad (24)$$

is a $2 \times N$ matrix and

$$E = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{xy} & \varepsilon_{yy} \end{pmatrix} \quad (25)$$

is a 2×2 covariance matrix of the populations' information-limiting correlations. This can also be expressed in block matrix form as

$$\Sigma = \Sigma_0 + UEU^T = \begin{pmatrix} X_0 & Z_0^T \\ Z_0 & Y_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_{xx}\mathbf{x}'\mathbf{x}'^T & \varepsilon_{xy}\mathbf{x}'\mathbf{y}'^T \\ \varepsilon_{xy}\mathbf{y}'\mathbf{x}'^T & \varepsilon_{yy}\mathbf{y}'\mathbf{y}'^T \end{pmatrix} \quad (26)$$

where X_0 and Y_0 are covariance matrices of \mathbf{r}_x and \mathbf{r}_y that do not limit information, and Z_0 is a cross-covariance between \mathbf{r}_x and \mathbf{r}_y that again does not limit information. The total linear Fisher information in both populations together can be expressed as

$$J = \mathbf{f}'^T \Sigma^{-1} \mathbf{f}' \approx \mathbf{1} E^{-1} \mathbf{1}^T = \frac{\varepsilon_{xx} + \varepsilon_{yy} - 2\varepsilon_{xy}}{\varepsilon_{xx}\varepsilon_{yy} - \varepsilon_{xy}^2} \quad (27)$$

where $\mathbf{1}=(1,1)$ and the approximation holds in the limit of many neurons with a Σ_0 that permits extensive information.

Under the condition that only population \mathbf{x} is decoded, yet is decoded optimally, then we have the unbiased weight vector $\mathbf{w}=(X_0^{-1}\mathbf{x}', 0)/J_x$ where $J_x \approx 1/\varepsilon_{xx}$ is the linear Fisher

information in population \mathbf{x} alone. We can now calculate the choice correlations in the \mathbf{y} population, according to $C_k = (\sum \mathbf{w})_k / \sqrt{\sum_{kk} \mathbf{w}^\top \sum \mathbf{w}}$, which yields

$$C_k = \frac{Z_o X_0^{-1} \mathbf{x}' / J_{0x} + \varepsilon_{xy} \mathbf{y}'}{\sqrt{\sum_{kk} (J_{0x}^{-1} + \varepsilon_{xx})}} \approx \frac{\varepsilon_{xy}}{\varepsilon_{xx}} \frac{\theta}{\theta_k} \quad (28)$$

The approximation holds in the limit of large populations (hundreds of neurons) where Σ_0 does not limit information. The choice probabilities in population \mathbf{y} can be larger than those in the optimally decoded population (Figure 5C) precisely when there is information-limiting noise in \mathbf{y} that is partially correlated with the information-limiting noise in \mathbf{x} , but with larger amplitude. Under such conditions, population \mathbf{y} alone has less information than \mathbf{x} , so reading out population \mathbf{x} alone can be only moderately suboptimal (Figure 5D).

There is a subtle technical point about how to reconcile Eq. 28, which shows that $C_k = \beta C_k^{\text{opt}}$ with β that can be greater than 1, with Eqs. 6 or 23, which state that

$C_k \approx \alpha C_k^{\text{opt}} + \sqrt{1 - \alpha} C_k^{\text{sub}}$ with a coefficient on C_k^{opt} that is less than or equal to 1. At first glance they seem inconsistent, because when $\beta > 1$ one might expect to see a term $\sqrt{1 - \beta} C_k^{\text{sub}}$ which cannot be valid. To resolve this, notice that the noise structure (Eq. 26) induces choice correlations C^{sub} that are proportional to C^{opt} for each population (with different proportionality constants ζ for each population that are determined by the correlations and decoder). As a result, the two terms in Eq. 6 can be combined into one:

$$\begin{aligned} C_k &\approx \alpha C_k^{\text{opt}} + \sqrt{1 - \alpha} C_k^{\text{sub}} \\ &\approx \alpha C_k^{\text{opt}} + \sqrt{1 - \alpha} \zeta C_k^{\text{opt}} \\ &\approx \beta C_k^{\text{opt}} \end{aligned} \quad (29)$$

which recovers Eq. 28. Thus while β measures the scale of choice correlations, α represents the fraction of uncertainty caused by information-limiting noise rather than suboptimal decoding and remains between 0 and 1. Even though they may be confounded by some measurements, α and β are conceptually distinct.

Tuning curves

Neural responses to multiple repetitions of eight azimuthal headings were recorded. Azimuthal tuning curves were fit by baseline-offset von Mises functions (Eq. 7). The parameters $\psi_k = \{b_k, a_k, \kappa_k, s_k\}$ were determined using maximum likelihood estimation assuming Poisson statistics for spike counts \mathbf{r} in a time window $t=1\text{sec}$ in response to eight uniformly spaced azimuth angles $s \in \{\pi/4, \dots, 2\pi\}$

$$\psi_k = \arg \max_{\psi} \sum_s (-f(s; \psi) \Delta t + r \log f(s; \psi) \Delta t) \quad (30)$$

In the maximization, the width parameter was subject to the constraint $\kappa < (8/\pi)^2$ corresponding to the requirement that the tuning curve width be no narrower than the spacing between the eight tested headings.

Fitting model predictions to data

Choice correlations were fit to the optimal and suboptimal predictors, Eqs. 3 and 2 respectively, using Bayesian multiple linear regression with heteroskedastic errors in variables ((Minka 1999), Supplemental Material). This method maximizes the likelihood L of the data after marginalizing over latent variables representing noiseless measurements. To assess the quality of the fits, we used the corrected Akaike Information Criterion (AICc, (Burnham 2004)), which is just -2 times the log-likelihood, $\log L$, with an additional penalty for having more free parameters: $AICc = -2\log L + 2p + 2p(p+1)/(n-p-1)$ for n data points and p parameters.

To compute these likelihoods we approximated measurement errors in C_k as normally distributed around the true values, with variances $\sigma_{C_k}^2$. Variances for experimentally measured choice correlations and threshold ratios (Eq. 3) were computed by 1000-fold bootstrapping. Variances for the suboptimal predictors (Eq. 2) were obtained from the second derivative of the log-likelihood used to fit the tuning curves (Supplemental Material). For model simulations, we generated normally-distributed measurement errors with variance

$\sigma_{C_k}^2 = \frac{1}{t-1} (1 - C_k^2)^2$ for t trials, which is the leading order expression for the variance of a correlation coefficient estimated from a finite number of bivariate gaussian samples (Kenney 1940).

Simulations

Simulated neural populations had 500 neurons. Tuning properties were sampled independently from experimentally recorded neurons. Median and central quartiles on those tuning parameters were approximately: $a=24\pm 20$ Hz; $b=0$ for 35% of cells, $b=13\pm 10$ Hz for the rest; $\kappa=1\pm 0.5$ radians⁻². Preferred stimuli s_k were drawn randomly from a uniform distribution over $[0, 2\pi)$.

Suboptimal correlation-blind decoders were factorial decoders unless otherwise specified.

For simulations with finite data (Figure 6), 50 neurons were sampled from the simulated population. Tuning curves were re-estimated from 10 responses to eight stimulus directions, as described above for real data. Simulated measurements of choice correlations were drawn from a gaussian distribution with the true mean C_k and variance given above by $\sigma_{C_k}^2$ for $t=30$ trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

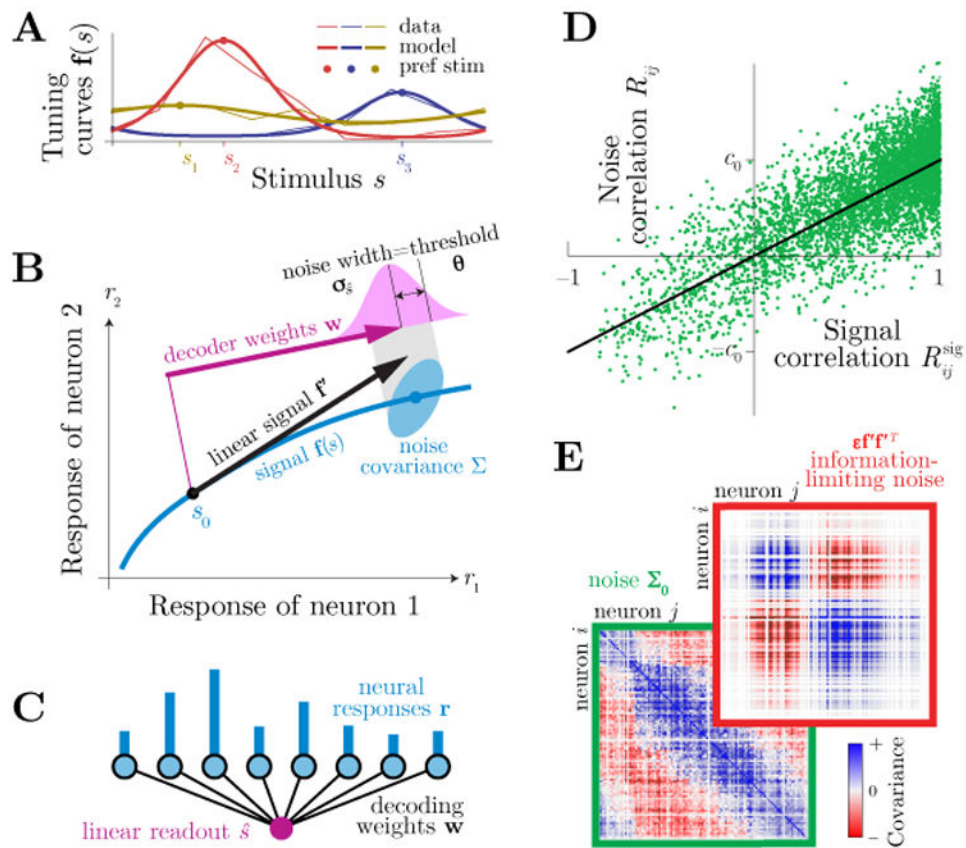
Thanks to Aihua Chen and Yong Gu for providing access to their neural recordings, and to Kaushik Lakshminarasimhan for helpful conversations. This work was supported by NIH grant T32DC009974 and a McNair Foundation grant to XP, NIH grant EY016178 to GCD, NIH grants EY017866 and DC004260 to DEA, and a James McDonnell Foundation grant and SNF grant 31003A-143707 to AP.

References

- Abbott LF, Dayan P. The effect of correlated variability on the accuracy of a population code. *Neural Comput.* 1999; 11(1):91–101. [PubMed: 9950724]
- Amos DE. Computation of modified Bessel functions and their ratios. *Math Comp.* 1974; 28:239–251.
- Berens P, Ecker AS, Cotton RJ, Ma WJ, Bethge M, Tolias AS. A fast and simple population code for orientation in primate V1. *J Neurosci.* 2012; 32(31):10618–10626. [PubMed: 22855811]
- Bondy A, Cumming B. Top down signals influence the distribution of noise correlations amongst sensory neurons. *Society for Neuroscience.* 2013
- Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci.* 1996; 13(1):87–100. [PubMed: 8730992]
- Britten KH, Shadlen MN, Newsome WT, Movshon JA. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci.* 1992; 12(12):4745–4765. [PubMed: 1464765]
- Burnham KPA, DR. Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods and Research.* 2004; 33:261–304.
- Chen A, DeAngelis GC, Angelaki DE. Representation of vestibular and visual cues to self-motion in ventral intraparietal cortex. *J Neurosci.* 2011; 31(33):12036–12052. [PubMed: 21849564]
- Chen A, DeAngelis GC, Angelaki DE. Functional specializations of the ventral intraparietal area for multisensory heading discrimination. *J Neurosci.* 2013; 33(8):3567–3581. [PubMed: 23426684]
- Cohen MR, Kohn A. Measuring and interpreting neuronal correlations. *Nat Neurosci.* 2011; 14(7):811–819. [PubMed: 21709677]
- Cohen MR, Newsome WT. Context-dependent changes in functional circuitry in visual area MT. *Neuron.* 2008; 60(1):162–173. [PubMed: 18940596]
- Cohen MR, Newsome WT. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *J Neurosci.* 2009; 29(20):6635–6648. [PubMed: 19458234]
- Drugowitsch J, DeAngelis GC, Klier EM, Angelaki DE, Pouget A. Optimal multisensory decision-making in a reaction-time task. *Elife.* 2014; 3
- Ecker AS, Berens P, Tolias AS, Bethge M. The effect of noise correlations in populations of diversely tuned neurons. *J Neurosci.* 2011; 31(40):14272–14283. [PubMed: 21976512]
- Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat Neurosci.* 2012; 15(1):146–154. [PubMed: 22101645]
- Földiák, P. The ‘Ideal Homunculus’: statistical inference from neural population responses. In: Eeckman, FH.; Bower, JM.; Norwell, MA., editors. *Computation in Neural Systems.* Kluwer Academic Publishers; 1993. p. 55-60.
- Ghose GM, Harrison IT. Temporal precision of neuronal information in a rapid perceptual judgment. *J Neurophysiol.* 2009; 101(3):1480–1493. [PubMed: 19109454]
- Graf AB, Kohn A, Jazayeri M, Movshon JA. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat Neurosci.* 2011; 14(2):239–245. [PubMed: 21217762]
- Green, DM.; Swets, JA. *Signal detection theory and psychophysics.* New York: Wiley; 1966.
- Gu Y, Angelaki DE, DeAngelis GC. Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience.* 2008; 11(10):1201–1210.
- Gu Y, DeAngelis GC, Angelaki DE. A functional link between area MSTd and heading perception based on vestibular signals. *Nat Neurosci.* 2007; 10(8):1038–1047. [PubMed: 17618278]

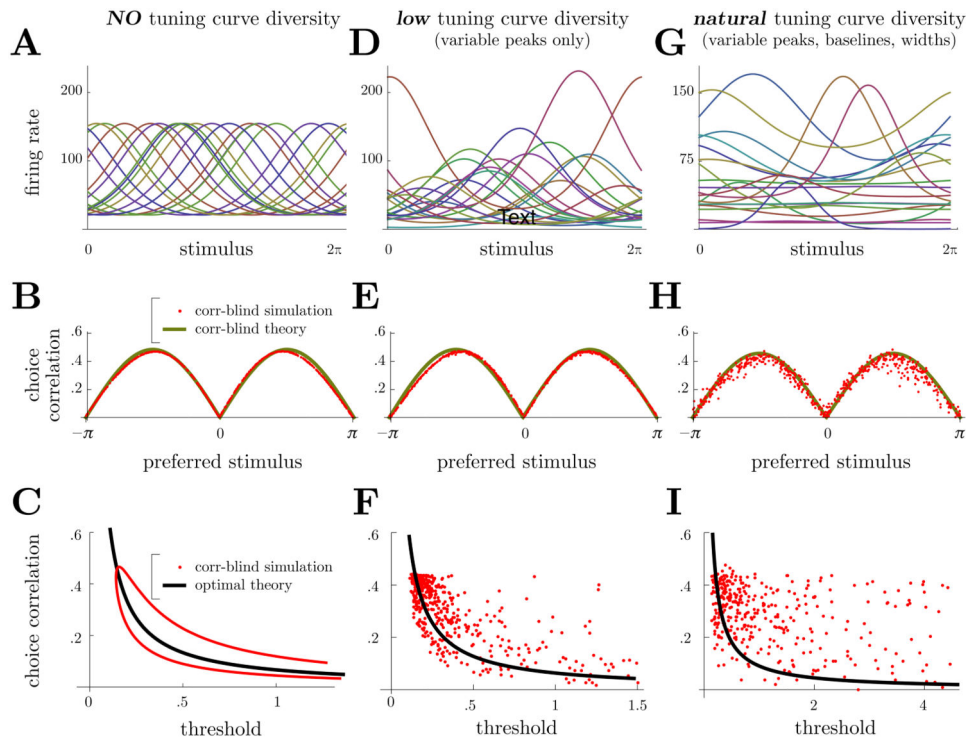
- Gu Y, Liu S, Fetsch CR, Yang Y, Fok S, Sunkara A, DeAngelis GC, Angelaki DE. Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*. 2011; 71(4):750–761. [PubMed: 21867889]
- Gu Y, Watkins PV, Angelaki DE, DeAngelis GC. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *J Neurosci*. 2006; 26(1):73–85. [PubMed: 16399674]
- Haefner RM, Gerwinn S, Macke JH, Bethge M. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nat Neurosci*. 2013; 16(2):235–242. [PubMed: 23313912]
- Kang I, Maunsell JH. Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. *J Neurophysiol*. 2012; 108(12):3403–3415. [PubMed: 22993262]
- Kenney, J. *Mathematics of Statistics, Part Two*. London: Chapman & Hall, Ltd; 1940.
- Klier, EM.; Liu, S.; Gu, Y.; DeAngelis, GC.; Angelaki, DE. Ventral intraparietal (VIP) inactivation does not affect multisensory heading perception. Annual Meeting of the Society for Neuroscience; San Diego. 2013.
- Lakshminarasimhan K, Liu S, Gu Y, Klier EM, DeAngelis GC, Pitkow X, Angelaki DE. Dissecting the contributions of MSTd and VIP to heading perception. *Cosyne abstract*. 2014
- Liu S, Gu Y, DeAngelis GC, Angelaki DE. Choice-related activity and correlated noise in subcortical vestibular neurons. *Nat Neurosci*. 2013; 16(1):89–97. [PubMed: 23178975]
- Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. *Nat Neurosci*. 2006; 9(11):1432–1438. [PubMed: 17057707]
- Meng H, Green AM, Dickman JD, Angelaki DE. Pursuit--vestibular interactions in brain stem neurons during rotation and translation. *J Neurophysiol*. 2005; 93(6):3418–3433. [PubMed: 15647394]
- Minka T. Linear regression with errors in both variables: A proper Bayesian approach. Unpublished manuscript. 1999
- Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A. Information-limiting correlations. *Nature Neuroscience*. 2014; 17:1410–1417.
- Nienborg H, Cumming B. Correlations between the activity of sensory neurons and behavior: how much do they tell us about a neuron's causality? *Current Opinion In Neurobiology*. 2010; 20(3):376–381. [PubMed: 20545019]
- Nienborg H, Cumming BG. Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. *Nat Neurosci*. 2007; 10(12):1608–1614. [PubMed: 17965712]
- Nienborg H, Cumming BG. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*. 2009; 459(7243):89–92. [PubMed: 19270683]
- Press, WHT.; Saul, A.; Vetterling, William T.; Flannery, Brian P. *Numerical recipes: the art of scientific computing*. Cambridge, UK; New York: Cambridge University Press; 2007.
- Purushothaman G, Bradley DC. Neural population code for fine perceptual decisions in area MT. *Nat Neurosci*. 2005; 8(1):99–106. [PubMed: 15608633]
- Salinas E, Abbott LF. Vector reconstruction from firing rates. *J Comput Neurosci*. 1994; 1(1-2):89–107. [PubMed: 8792227]
- Sanger TD. Probability density estimation for the interpretation of neural population codes. *Journal Of Neurophysiology*. 1996; 76(4):2790–2793. [PubMed: 8899646]
- Schwartzman A. Lognormal Distributions and Geometric Averages of Positive Definite Matrices. *arXiv*. 2014 stat.ME (1407.6383).
- Shamir M, Sompolinsky H. Implications of neuronal diversity on population coding. *Neural Comput*. 2006; 18(8):1951–1986. [PubMed: 16771659]
- Sompolinsky H, Yoon H, Kang K, Shamir M. Population coding in neuronal systems with correlated noise. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2001; 64(5 Pt 1):051904. [PubMed: 11735965]
- Tolhurst DJ, Movshon JA, Dean AF. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res*. 1983; 23(8):775–785. [PubMed: 6623937]
- Uka T, De Angelis GC. Contribution of area MT to stereoscopic depth perception: choice-related response modulations reflect task strategy. *Neuron*. 2004; 42(2):297–310. [PubMed: 15091344]

- Wohrer A, Machens C. Percept formation from neural populations in sensory decision-making tasks. arXiv. 2013 1303.1939.
- Wu S, Nakahara H, Amari S. Population coding with correlation and an unfaithful model. *Neural Comput.* 2001; 13(4):775–797. [PubMed: 11255569]
- Yates J, Katz L, Park IM, Pillow J, Huk AC. Dissociated functional significance of choice-related activity across the primate dorsal stream. *Cosyne abstract.* 2014
- Zohary E, Shadlen MN, Newsome WT. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature.* 1994; 370(6485):140–143. [PubMed: 8022482]

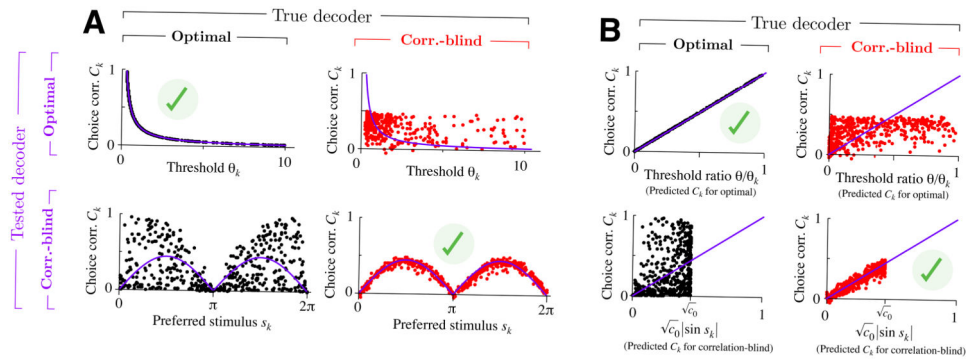
**Figure 1.**

Model for neural responses and decoding. (A) Three example tuning curves $f(s)$ show the mean neural responses to a stimulus s (thin lines). These tuning curves are modeled by von Mises functions (thick curves) with parameters including the preferred stimulus s_k (dots). (B) As the stimulus s varies, the mean activity of all neurons traces out a curve $f(s)$ (blue) through the N -dimensional space of neural responses. For fine local discriminations, the relevant signal direction lies along the tangent f' (black). Neural responses are decoded by projection onto a readout direction w (magenta). The amplitude of w is scaled to give an unbiased estimate \hat{s} of the stimulus, so that a unit change in stimulus generates a unit change in the estimate. Trial-to-trial variability is expressed as multivariate gaussian noise, with covariance Σ (ellipse). The projection of this noise along the decoding direction (pink gaussian) has standard deviation $\sigma_{\hat{s}}$, which we define as the population threshold θ . Although we illustrate responses for only 2 neurons here, these relationships generalize to high-dimensional response spaces. (C) Linear decoding projects the neural responses, both noise and signal, onto a particular direction w to obtain an estimate \hat{s} of the stimulus. (D) Noise correlation coefficients R_{ij} between distinct neurons i and j are modeled as being proportional on average to the signal correlations R_{ij}^{sig} , with proportionality c_0 . This means that neurons with similar tuning tend to have more correlated fluctuations. (E) We distinguish two components to the noise covariance Σ : information-limiting correlations that lie along the signal direction f' and thus have covariance ff'^T (matrix boxed in red), and the remaining noise with covariance Σ_0 (matrix boxed in green). The two forms of noise have

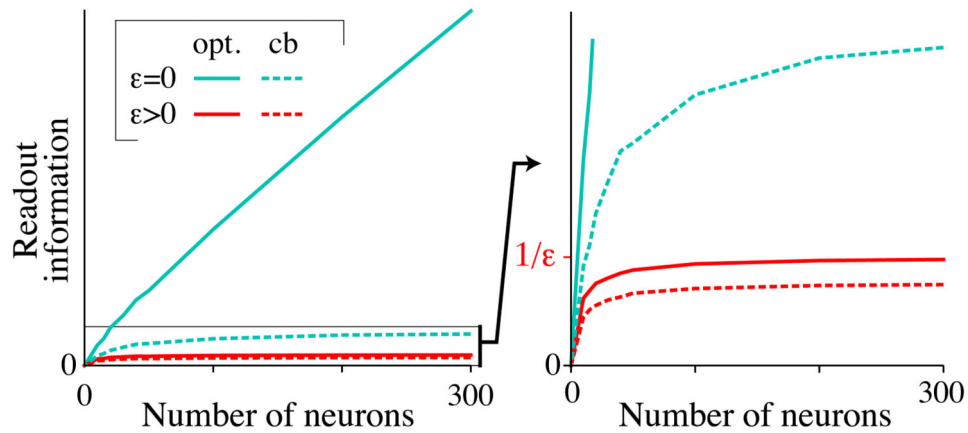
distinct structures that are apparent in the covariance matrices. The striations in both matrices reflect the heterogeneous tuning curve amplitudes.

**Figure 2.**

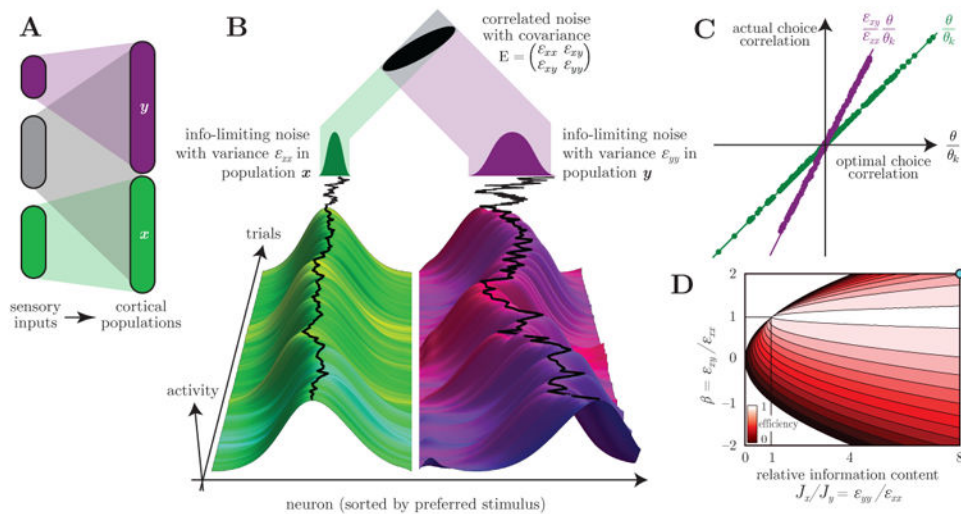
As neural heterogeneity increases, the choice correlations predicted for optimal and suboptimal decoding become increasingly distinct. With homogeneous tuning curves (**A**), the preferred heading and neural threshold are closely related, as the neural threshold is smallest when the stimuli to be discriminated lie along the steep flank of the tuning curve. (**B**) The choice correlations for a simulated population read out using a suboptimal correlation-blind decoder (red) are well described by the predicted sinusoid (green). (**C**) However, the optimal decoding predictor (black) is also able to fit those same choice correlations (red) reasonably well, especially in the presence of typical measurement noise. This is because choice correlations for the optimal decoder are inversely related to the neural discrimination threshold, which, again, has a direct relationship to the preferred stimulus for a homogeneous population. When there is neural diversity only in tuning curve amplitude (**D**), similar though more variable relations hold (**E,F**). With the large diversity in amplitude, baseline, and width observed in measured tuning curves (**G**), choice correlations for a correlation-blind decoder remain well described by the correlation-blind predictor. This is true even with correlations that do not have noise correlations that are precisely proportional to the signal correlations (Eq. 8) but rather exhibit some diversity around that trend (Eq. 10) (**H**). In contrast, with all of this heterogeneity, those suboptimal choice correlations become poorly fit by the optimal decoding predictor (**I**).

**Figure 3.**

Optimal and suboptimal, correlation-blind decoding of simulated heterogeneous neural populations can be readily distinguished by their predicted patterns of choice correlations. Top and bottom rows show the choice probabilities of optimally and suboptimally decoded neural populations, plotted against predictions of correlation-blind (Eq. 2) and optimal (Eq. 3) decoders. In (A) simulated choice correlations are plotted against neural threshold θ_k and preferred stimulus s_k , quantities with which they should have nonlinear relationships. In (B) the same simulated choice correlations are plotted against θ/θ_0 and $\sqrt{c_0} |\sin s_k|$, where a linear relationship should hold for optimal and suboptimal decoding, respectively. Green checks indicate good agreement when the prediction matches the true decoder. Note that the bottom-left panel of (B) provides a direct comparison of the two predictions in the case of heterogeneous tuning curves, since the horizontal axis indicates choice correlations for the correlation-blind decoder while the vertical axis gives choice correlations for the optimal decoder (which is the true decoder for that simulation).

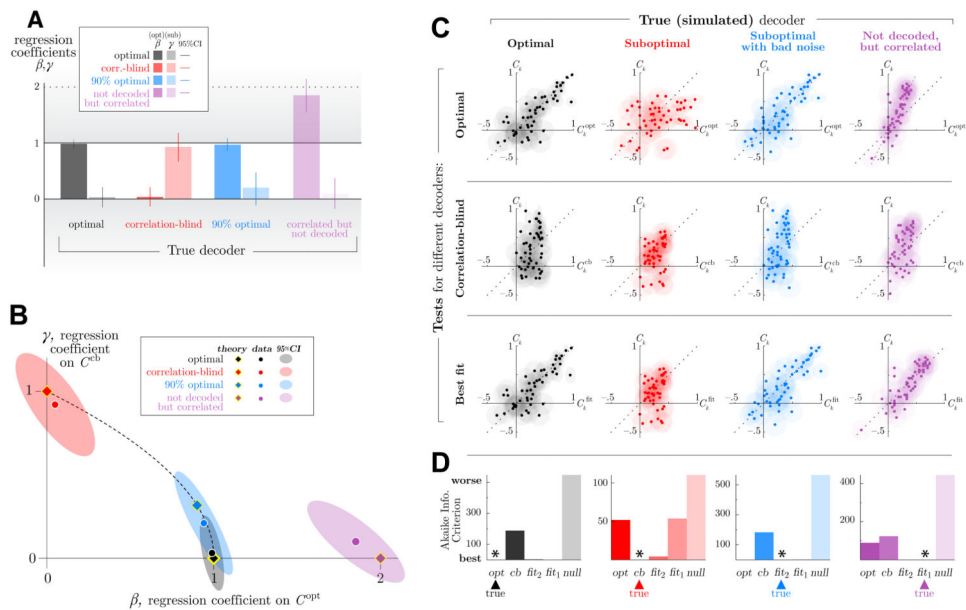
**Figure 4.**

Effects of information-limiting noise and correlation-blind decoding on the information extracted by a linear estimator. Optimal decoding of a neural population without differential correlations yields extensive information that increases without bound as more neurons are added (solid cyan). Suboptimal, correlation-blind (cb) decoding of the same population is only able to extract a limited amount of information even with an infinite number of neurons (dashed cyan). In the presence of differential correlations with variance ϵ (Eq. 4), however, information saturates to a finite value of $1/\epsilon$ even for large populations that are decoded optimally (solid red). The suboptimal (cb) decoder does not perform much worse than the optimal decoder in this case, extracting information (dashed red) which is not much smaller than $1/\epsilon$. The second panel is a vertically expanded view of the first panel.

**Figure 5.**

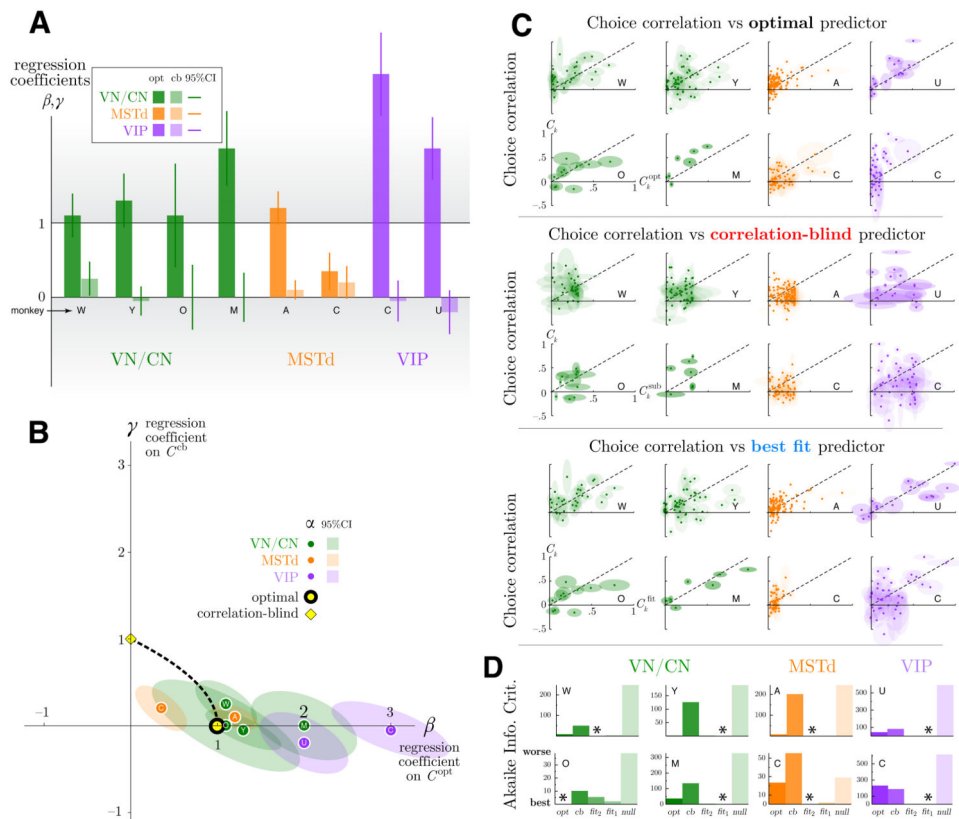
Choice correlations for two populations, x (green) and y (purple), with correlated information-limiting noise. **(A)** Schematic of one way that the two large cortical populations can inherit this form of noise by receiving some shared sensory input (grey ellipse) in addition to their own private sensory inputs. **(B)** Illustration of activity in both cortical populations as a consequence of correlated information-limiting noise. Information-limiting noise causes the neural activity in each population to fluctuate from trial to trial (greenish and purplish surfaces). As shown here, these fluctuations are visualized most readily for homogeneous neural populations with pure information-limiting noise —fluctuations that look exactly as if the stimulus itself had shifted (black curves). Over many trials, this variability has a distribution (shown above the neural activity) whose width is determined by ϵ and inversely related to the information content. If the information-limiting noise is identical between the two populations, then the fluctuations will have the same extent, but if the populations have at least partially distinct sources of information, then the information-limiting noise in each will be partially correlated (black ellipse). This example has a variance of $\epsilon_{xx}=1$ in the green population, and a higher variance $\epsilon_{yy}=8$ in the purple

population, with covariance $E \propto \begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix}$. For these parameters, half of the variance in the purple population copies the fluctuations in the green population, but with twice the size. **(C)** If only population x is decoded — and decoded optimally — then its choice correlations are given exactly by Eq. (3) (green line). Choice correlations for the non-decoded population y (purple) are proportional to those optimal predictors, but with a proportionality $\epsilon_{xy}/\epsilon_{xx}$ (Eq. 28) that can be greater or less than 1. This analytical result (solid lines) is supported by simulations of neurons in two populations (filled symbols) with noise covariance described by Eq. (26). **(D)** Contour map of the decoder efficiency, which is the fraction of the total information in the population that is actually extracted by the decoder. For a suboptimal decoder that extracts essentially all of the information J_x in population x but uses none of the responses from population y , the efficiency can range from highly suboptimal ($J_x/J \approx 0$) to nearly optimal ($J_x/J \approx 1$) depending on the covariance structure E . For the covariance depicted in **B** and **C**, this decoder achieves an efficiency of 80% (cyan dot).

**Figure 6.**

Inferring decoding quality from simulations with realistically limited amounts of data. We use numbers of neurons (50) and trials (30 repetitions of each stimulus) that are comparable to those obtained in our experiments. Choice correlations are fit to predictions from optimal (Eq. 3) and correlation-blind (Eq. 2) decoding, and the resultant regression coefficients (β , γ) are plotted separately as a pair of bars in (A) and jointly as a point in (B), indicating how much of variance in behavior can be explained by optimal or correlation-blind decoding. In particular, $\beta=1$ when decoding is optimal, with its quality limited by noise rather than by suboptimal decoding. Four example populations and decoders are shown: optimal (black), correlation-blind (red), correlation-blind with information-limiting noise (blue), and an undecoded population that is correlated with an optimally decoded population. For each example, the recovered coefficients (filled circles) for the decoded populations fall within 95% confidence intervals (shaded ellipses in B) of the parameters expected for the true decoder (filled diamonds). Thus, the optimal decoder is revealed as optimal (black), and the correlation-blind decoder is revealed as clearly suboptimal for a population with extensive information (red) but nearly optimal for a population with limited information (blue). The parametric curve $(\alpha, \sqrt{1-\alpha})$ (dashed) shows coefficients expected for models with a fraction α of total uncertainty caused by information-limiting noise. For the undecoded population, coefficients fall near the theoretical values (2,0) used in the simulation (purple), correctly implying that the population cannot account for the quality of the decoded output. (C) Choice correlations for individual simulated neurons are plotted against those predicted for optimal decoding ($C_k^{opt} = \theta / \theta_k$, the psychophysical/neural threshold ratio, Eq. 3), correlation-blind decoding ($C_k^{cb} = \sqrt{c_0} |\sin s_k|$, Eq. 2), and correlation-blind decoding combined with substantial information-limiting noise. For this last decoder, the information-limiting noise is responsible for a fraction β of the behavioral variance, where β is determined by fitting ($C_k^{fit} \approx \beta C_k^{opt} + \gamma C_k^{cb}$). Shaded ellipses again represent 95% confidence intervals for each data point. (D) Relative quality of statistical models, assessed

using the corrected Akaike Information Criterion (AICc, see Methods). Smaller AICc are much better, because the relative probability of model i given the data depends exponentially on the AICc according to $P_i \propto e^{-\frac{1}{2}AICc_i}$. We compare AICc for five models that predict choice correlations: purely optimal decoding (*opt*), a correlation-blind decoder that ignores correlated noise (*cb*), a two-parameter fit based on a weighted sum of the *opt* and *cb* predictors (*fit₂*), a one-parameter fit using only a scaled version of the *opt* predictor (*fit₁*), and a null model attributing all variability to random errors (*null*). In each example shown here, the model with the lowest AICc (asterisk) is in fact the correct one (triangle marked ‘true’). Since *fit₁* is a special case of *fit₂*, and *opt* is a special case of *fit₁*, then if the best explanation of the choice correlation data is indeed the optimal predictor, then the *fit₁* and *fit₂* models also provide good explanations, although they are penalized for having extra parameters (black, blue). On the other hand, the undecoded population is not optimally decoded but its choice correlations can be explained as a scaled version of the optimal choice correlations, and is thus best explained by *fit₁* or *fit₂* (purple).

**Figure 7.**

Inferring the decoding quality from experimental data, plotted as in Figure 6. Neural populations were recorded from VN/CN (green), MSTd (orange), and VIP (purple). Choice correlations are fit to predictions from optimal (Eq. 3) and correlation-blind (Eq. 2) decoding, and the resultant regression coefficients (β, γ) are plotted separately as a pair of bars in (A) and jointly as a point in (B). Each datum combines all recordings from one brain area in one monkey, whose initials are shown below the bars or inside the point. (C) Scatter plots of measured choice correlations for individual neurons are plotted against choice correlations predicted from the optimal (*opt*) and correlation-blind (*cb*) models, as well as the best fitting linear combination of those two predictors (*fit₂*). Each plot has the initial of the monkey from whom the data was recorded. The plots are arranged in pairs of rows only for compactness. (D) The quality of these three models, plus two additional models *fit₁* and *null*, was assessed using the AICc, as described for Fig. 6. Based on these measures of model quality, data from the heading-discrimination task are largely consistent with animals using VN/CN and MSTd information near-optimally, while neglecting less informative responses in VIP.